DOCUMENT RESUME

ED 266 628 FL 015 442

AUTHOR	Kohonen, Viljo, Ed.; And Others
TITLE	Practice and Problems in Language Testing 8. Papers presented at the International Language Testing Symposium of the Interuniversitare Sprachtestgruppe (8th, Tampere, Finland, November 17-18, 1984). AFinla Series No. 40.
INSTITUTION	Finnish Association for Applied Linguistics, Helsinki.
REPORT NO	ISBN-951-9388-27-3
PUB DATE	85
NOTE	216p.; For individual papers, see FL 015 443-459.
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Guides - Classroom Use - Guides (For Teachers) (052) -- Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01/PC09 Plus Postage.
DESCRIPTORS	Adolescents; Communicative Competence (Languages); Cultural Awareness; *English (Second Language); Error Patterns; *Evaluation Criteria; Foreign Countries; Holistic Approach; Intercultural Communication; Language Laboratories; *Language Proficiency; *Language Tests; Microcomputers; Reading Comprehension; Secondary Education; Second Language Learning; Teacher Education; Test Validity; Vocabulary
IDENTIFIERS	Czechoslovakia; West Germany

ABSTRACT

A collection of papers on language testing addresses these topics: language tests seen from a communicative viewpoint; microcomputers and testing; productive second language skills; developing teachers' assessment skills; content validity of holistic tests; statistical regression in group conversation tests; reading comprehension; unique articulatory and auditory dimensions for English as a second language data; foreign language needs in Czechoslovakia; language test dimensionality; adolescent native language proficiency assessment; assessment at the end of secondary school; estimating the quantity of students' second language vocabulary; the language laboratory; informal assessment; cultural knowledge as part of language skills in cross-cultural communication; and classroom English for German schools. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

BEST COPY AVAILABLE

SUOMEN SOVELTAVAN KIELITIETEEN YHDISTYKSEN
(AFinLA) JULKAISUJA

PUBLICATIONS DE L'ASSOCIATION FINLANDAISE
DE LINGUISTIQUE APPLIQUÉE (AFinLA)

No. 40

ED266628

Practice and Problems in Language Testing 8

edited by

Viljo Kohonen,
Hilkka von Essen and Christine Klein-Braley

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Tampere 1985

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J Tommola

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

FL 015442

Companion Volume:

Kohonen, Viljo and Antti J. Pitkänen (eds.). 1985. Language Testing in School. AFinLA Yearbook 1985. AFinLA Series 41. Tampere.

FOREWORD
CONTENTS

Kalevi Pohjala, Opening of the Heinola Seminar 5
Sauli Takala, Criterion-referenced testing in foreign language teaching 9
Christine Klein-Braley, Tests of reduced redundancy - theory 33
Ulrich Raatz, Tests of reduced redundancy - the C-test, a practical example 49
Lars Sigfred Evensen, Testing language skills in the Norwegian comprehensive school 63
John H.A.L. de Jong, Effect of mixing ability groups on ability levels attained 79
Viljo Kohonen, Testing communication skills in mixed-ability groups 91
Torsten Lindblad, On the whys and hows of language testing 109
Jan Prucha, Current trends in FL testing in Czechoslovakia and the Soviet Union 117
Elana Shohamy, Thea Reves and Yael Bejarano, From testing research to educational policy: introducing a new comprehensive test of oral proficiency 133
Bernd Voss, Language testing in German schools - theory and practice 145
Douglas K. Stevenson, The teaching of testing: reliability, validity, and all that stuff 165
Anneli Kauppinen, Äidinkielen taidon mittaaminen lukion päättövaiheessa 181
Heljä Kauhanen, Kommentti Anneli Kauppisen puheenvuoroon äidinkielen taidon mittaamisesta lukion päättövaiheessa 189
Kyllikki Keravuori, Koululuokan diskurssin piirteitä 195
Matti Luukkainen, Wissen und Können im Erwerb der dritten und vierten Fremdsprache 209
Antti J. Pitkänen, Om språkliga drag och feltyper vid fri skriftlig framställning i främmande språk 227
Pauli Roinila, Tekstilajit ja kirjoitetun tekstin muotoutuminen 239
Vuokko Vuori-Barry, Kirjoittamisen opettamisesta lukiossa (A-englanti) 251

© The Finnish Association for Applied Linguistics (AFinLA) and the authors

ISBN 951 - 9388 - 27 - 3
ISSN 0781 - 0318

FOREWORD

The present volume contains most of the papers presented at the Eighth International Language Testing Symposium of the Interuniversitäre Sprachtestgruppe (IUS) at the University of Tampere, November 17-18, 1984. The symposium was immediately followed by a seminar on "Language Testing in School" at the Heinola Course Centre, November 19-21. In connection with the IUS Symposium, there was a parallel symposium of the Finnish Association for Applied Linguistics (AFinLA) at the University of Tampere. The central theme of all the three occasions was **"Language - Knowledge and Skill"**, a theme emphasizing the need to examine the current communicative goal orientation from the point of view of both linguistic competence and the ability - and the courage and willingness - to put that competence into pragmatically meaningful and socially acceptable use.

An important aim of the symposiums was to exchange information about current thinking and practices in various national contexts of language teaching; we have a great deal to learn from each other. An attempt was therefore made to obtain a wide international basis for sharing ideas. We wish to express our gratitude to the following institutions for their generous financial support: the America Center, the British Council, the Goethe-Institut, the Nordic Language and Information Centre, the Tampere University Foundation, and the Finnish Ministry of Education. Our thanks are also due to the University of Tampere for providing the facilities for the Tampere symposiums, and to the Heinola Course Centre for the warm hospitality at Heinola. Without all this support, the organization of the symposiums would not have been possible.

To disseminate information further to those who were not able to join us in Finland, most of the papers presented at Tampere and Heinola are published in two companion volumes in the AFinLA Series: the present one, and the AFinLA 1985 Yearbook, containing the papers read at the AFinLA Symposium and at the Heinola Seminar; some of the papers will appear elsewhere in due course. We are grateful to Riitta Jaatinen and Jaana Rantanen for their help at the Tampere symposiums, to Ian Gurney for checking the English in some of the papers, and to Jouko Jousea and Riitta Könönen for their help with word processing.

The Editors

CONTENTS

Lyle F. Bachman, An Examination of some language proficiency tests from a communicative viewpoint..... 5
Terry Culhane, Microcomputers and language testing 21
Peter Dickson, Evaluation of the productive skills in the second language the work of the assessment of performance unit (APU)..... 27
Mary C. Kilborn, Developing classroom teachers' assessment skills 41
Christine Klein-Braley, Reflections on the content validity of holistic tests 53
Erkki P. Liski and Simo Puntanen, A percentile regression model for the number of errors in group conversation tests 63
Madeline Lutjeharms, The development of reading comprehension in FL teaching - combining knowledge and skills 77
Kyle Perkins, Sheila R. Brutten, and Paul J. Angelis, Evidence for unique articulatory and auditory dimensions in ESL data 89
Jan Prucha, Foreign language needs: theory and empirical evidence in Czechoslovakia..... 105
Ulrich Raatz, Investigating dimensionality of language tests - a new solution to an old problem 123
Klaas Schreuder, Formative tests of speaking proficiency in the mothertongue for 14-year olds..... 137
Elana Shohamy, An alternative model for assessing language proficiency at the end of secondary school..... 149
Sauli Takala, Estimating students' vocabulary sizes in foreign language teaching 157
Robert Vanderplank, Evaluating the language laboratory in practice..... 167
Jan van Weeren, Language testing: making a decision on the basis of observed behaviour and 'tapped' knowledge..... 179
Pertti Widén, Cultural knowledge as part of language skills in cross-cultural communication 191
Bernd Voss, Classroom English for German schools 201

AN EXAMINATION OF SOME LANGUAGE PROFICIENCY TESTS FROM A COMMUNICATIVE VIEWPOINT

Lyle F. Bachman (University of Illinois at Urbana-Champaign)

1 Introduction

The theme of this conference, "Language = Knowledge and Skill," is particularly well-chosen, since it represents, I believe, a recognition that language proficiency involves **both** knowledge, or competence, and skill in implementing, or executing that competence. Skills and components models such as those proposed by Lado (1961) and Carroll (1961) distinguished skills (listening, speaking, reading, and writing) from components of knowledge (grammar, vocabulary, phonology/graphology), but did not indicate how these were related. It was not clear whether the skills were simply manifestations of the knowledge components in different modalities and channels, or whether they were qualitatively different in some other ways. For example, does reading differ from writing only in that it involves reception rather than production? If that were so, how can we account for the fact that quite competent and skillful readers are not always skillful writers? Chomsky's model (1965), with its distinction between competence and performance, permitted us to distinguish random "noise" from language proficiency, but in so doing limited language proficiency solely to competence. And neither of these models recognized the full context of language use--the contexts of discourse and situation. Halliday's framework (1976), with its focus on functions, both illocutionary and textual, clearly recognizes the context of discourse, but again is limited to competence. Finally, although Hymes' (1972) notion of sociolinguistic appropriateness recognizes the interaction between language use and the context of situation, it does not address the distinction between competence and skill.

Recent frameworks of communicative competence (Munby, 1978; Widdowson, 1978; Canale & Swain, 1980; Savignon, 1983), provide a much more inclusive description of the the knowledge required to use language, in that they incorporate linguistic competence, discourse competence, and sociolinguistic competence. All of these frameworks comprise what might be called descriptive rather than working models in that they focus on competence and either explicitly or implicitly ignore the implementation of that competence in language use. A more cognitive approach to language use has been taken in working models of language processing such as those proposed by Faerch and Kaspar (1983) and Bialystok and Ryan (forthcoming). But while such models distinguish planning from execution and characterize varying degrees of cognitive control in language processing, they do not specify how language competencies are distinguished from language skills.

At this point it may be useful to discuss what evidence there is for a distinction between competence and skill. The first kind of evidence lies in the differential implementation of competence in different skill modes. It is well-known, for example, that comprehension generally precedes production, both in first and in second language acquisition. If the requisite competencies are

present for comprehension, what causes this implementation in production to lag behind? Likewise with differences in channel. Fluent reading does not necessarily imply fluent listening, or vice-versa. Another kind of evidence can be found in the literature on communication strategies, which deals with the ways in which individuals attempt to achieve some communication goal in a given language despite inadequate communicative competence in that language. In such cases the individual may be able to compensate for inadequate competence by the implementation of various strategies. Thus, the most skillful communicator may not necessarily be the individual with the greatest competence.

To assure you that I am not merely juggling terms, I will outline a theoretical framework that distinguishes the knowledge, or competence aspects of language proficiency from the skills aspects, and that also addresses the factors in the language testing situation that affect performance on language tests. I believe this model is of use in examining the construct validity of tests of language proficiency and in clarifying some terms and concepts that have been used to describe various aspects of language proficiency. I will then examine, within this framework, the nature of the performance tasks and language competencies required by some widely used testing procedures. Finally, I will discuss the implications that this examination suggests for psychometric theory and for language testing research.

2 A framework for describing performance on tests of language proficiency

Adrian Palmer and I have proposed a framework for describing the different factors that affect performance on language tests (Bachman and Palmer, forthcoming). This framework includes four types of factors: language trait factors, skill factors, method factors, and random factors. Language trait factors are those competencies or mental abilities that are specific to language use, and are of two main types: organizational competence and pragmatic competence. **Organizational competence,** which includes grammatical and discourse competence, pertains to the formal characteristics of language usage. **Pragmatic competence,** which includes illocutionary and sociolinguistic competence, pertains to the functional and social characteristics of language use.

Skill factors are those general characteristics of the individual that affect test performance. These consist of 1) psycho-physiological skills, which are distinguished in terms of mode (productive/receptive) and channel (aural-oral/visual), 2) forms of representation (conscious/subconscious, analyzed/unanalyzed, pre-fabricated routines, rules), which determine the extent to which language competencies are available for use, and 3) strategic competence, which consists of a set of general abilities that affect how language competencies are implemented for maximum effectiveness in processing information.

Method factors are those characteristics of the test method that affect performance. These factors consist of 1) the type of language use situation (reciprocal/nonreciprocal), 2) the amount of context (embedded/reduced), 3) the distribution of information (compact/diffuse), 4) the type of information presented (concrete/abstract), and 5) the type and degree of restrictions on language performance; these include restrictions on the organization of discourse, the language use situation, propositional content, illocutionary force, forms, participants, mode, channel and time/length. "Communicative" testing methods might be characterized as those involving relatively unrestricted, appropriately contextualized language performance, while "non-communicative" testing methods involve only artificially restricted, inappropriately contextualized language performance.

Finally, random factors consist of 1) cognitive and affective characteristics of the individual, such as field dependence/independence, inhibition, tolerance/intolerance of ambiguity, and motivation, 2) interactions among specific combinations of trait, skill and method factors, and 3) measurement error.

One application of this framework has been in the definition of terms that have been used to refer to various aspects of language proficiency (Bachman and Palmer, 1984). Linguistic competence may be defined as consisting of the trait factors of grammatical competence: syntax, morphology, and phonology/graphology, while communicative competence consists of linguistic competence plus the trait factors of discourse competence, illocutionary competence, and sociolinguistic competence.

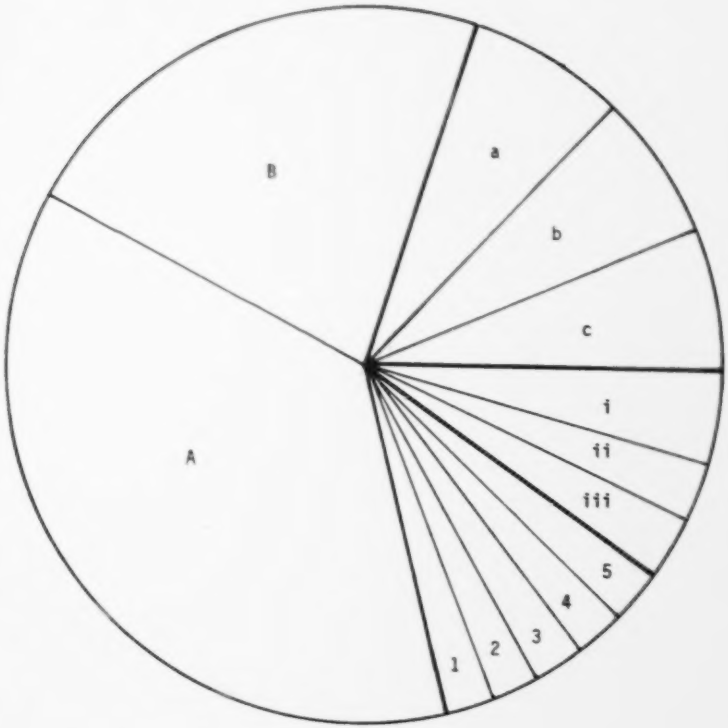
Language skills (listening, speaking, reading, writing) consist of trait and skill factors. Linguistic performance consists of the manifestation of linguistic competence and skill factors in artificially restricted and inappropriately contextualized test situations (methods). Communicative performance consists of the manifestation of communicative competence and skill factors in relatively unrestricted and appropriately contextualized test situations (methods).

A measure of linguistic performance includes that portion of a test score attributable to linguistic competence, skill factors, artificially restricted and inappropriately contextualized method factors, and random factors. A measure of communicative performance includes that portion of a test score attributable to communicative competence, skill factors, relatively unrestricted and appropriately contextualized method factors, and random factors.

This framework may also by useful in explaining sources of variation in performance on tests, as illustrated in the figure below.

I. Trait Factors:
   A. Organizational Competence
   B. Pragmatic Competence

II. Skill Factors:
   a. Psychophysiological
   b. Forms of Representation
   c. Strategic Competence



III. Method Factors:

1. Language Use Situation
2. Amount of Context
3. Distribution of Information
4. Type of Information
5. Restrictions on Language Performance

IV. Random Factors:

i. Cognitive and Affective Qualities
ii. Interactions among other Factors
iii. Measurement Error

Sources of Variation in Language Test Scores

The relative contribution of trait, skill, method and random factors to test performance will, of course, vary from test to test and from individual to individual. For example, Bachman and Palmer (1982) found that a multiple-choice test of grammatical competence loaded much more heavily on the method factor than did multiple-choice tests of either pragmatic or sociolinguistic competence. The effect of the task on test performance has generally been dealt with psychometrically as systematic error variance associated with the test method (Campbell & Fiske, 1959). The framework described here specifies in more detail the factors that comprise test method and at the same time recognizes the relationship between the demands set by the task and context of the test and the competencies required to successfully meet these demands.

Finally, this framework may be useful in clarifying some misconceptions regarding the terms "direct" and "indirect" as they have been applied to language tests. The term "direct test" is often used to refer to a test method in which performance resembles "actual" or "normal" language performance, while an "indirect test" is one in which test performance is perceived as somehow different from "actual" or "normal" performance. Thus, writing samples and oral interviews are referred to as "direct" tests, since they presumably involve the use of the skills being tested. By extension, such tests are often regarded, virtually without question, as construct valid and therefore as legitimate criteria for the validation of "indirect" tests.

There are two problems with this, however. First, we have no definition of "actual" or "normal" language use that is precise enough for us to determine the extent to which performance on a given test is similar to such language use. Indeed, the framework suggested here may at best permit us only to distinguish relatively "communicative" from relatively "non-communicative" language performance. A more serious problem, however, is that the use of the term "direct" confuses the behavioral manifestation of a trait or competence for the construct itself. As with all mental measures, language tests are indirect indicators of the underlying traits in which we are interested. The framework presented above captures this distinction by recognizing that there are factors in addition to trait factors that affect performance on all language tests, whether these require recognition of the correct alternative in a multiple-choice format or the writing of an essay.

3 An examination of some tests of language proficiency

In examining proficiency tests as measures of communicative performance, there are two questions that should be addressed. First, to what extent do the tasks required on the test involve communicative language performance? Second, to what extent does the test assess communicative competencies?

3.1 Multiple-choice tests

The multiple-choice test is one of the the most widely-used techniques for testing language proficiency in the world. Such tests typically include parts aimed at measuring at least some of the

following skills or components: 1) listening comprehension, 2) structure, and 3) reading comprehension. The language performance tasks on such tests are almost always restricted to non-reciprocal situations, in which there is no potential for feed-back or negotiation of meaning. The amount of context generally varies considerably from part to part, as does the distribution and type of information. The format of these tests generally restrict the mode of performance to reception. Finally, there are obvious restrictions on time and length.

3.1.1 Multiple-choice tests of listening comprehension

Multiple-choice tests of listening comprehension typically include tasks such as 1) listening to a sentence and then identifying the correct paraphrase from several choices, 2) listening to short dialogues and then finding the correct choice to a question about the dialogue, and 3) listening to a short talk and then answering comprehension questions based on that talk. The propositional content of this type of test is typically restricted to academic topics, while the illocutionary acts are typically ideational and manipulative. In the paraphrase item type two basic tasks are required: 1) comprehending a single spoken sentence (stem) and 2) recognizing the correct paraphrase of this sentence (key) from among four written sentences. The majority of the items of this type require only grammatical competence for successful completion. Most depend primarily on the knowledge of lexical signification, or of the propositional content expressed by syntactic structure. Furthermore, these items can be regarded as context-reduced, in that they are generally unconnected with each other and their references are to fictitious persons, objects, place and actions. In general, the task of recognizing paraphrases is an extremely artificial one and requires virtually no communicative performance, in that this task focuses exclusively on propositional signification. In addition, the lack of context renders this item type highly artificial.

In the lecture, or short talk test type there is generally a variety of discourse organization, including generalization and development, as well as a variety of illocutionary acts and linguistic forms. The basic tasks in this test type are 1) comprehending a spoken discourse and 2) answering direct information questions based on that discourse. In this type of test, the extent of the discourse is much more substantial than that in the paraphrase type. The context is also much more extensive. Unfortunately, however, the lectures are frequently highly artificial, in that they sound like "read" presentations and fail to include the kinds of hesitations and restatements that characterize oral discourse. There is little challenge to the test taker to interact with the text and consequently little opportunity for authentic language use.

3.1.2 Multiple-choice tests of structure

The language performance on this type of test is typically restricted entirely to single sentences, and thus has little potential for involving discourse. There is generally a variety of propositional content, while the illocutionary force is typically restricted to the

ideational function and the form to declarative sentences. In this test type the basic task is to recognize the syntactic form that will correctly complete an incomplete statement. The items in this type of test are generally context-reduced, in that they represent isolated propositions, although there is generally some attempt to contextualize them. Frequently, however, this context is totally irrelevant to the task posed by the item. Consider the following item, for example:

The first ornithischia appeared on the Earth _____ the early Mesozoic era, some 200 million years ago.

(A) when
(B) or
(C) and
(D) during

This statement, if found in an authentic discourse, would presuppose that the reader is familiar with the terms "ornithischia" and "Mesczoic era". In the test context, however, this information is irrelevant to the syntactic structure that requires the preposition "during". If the test taker attempts to process this sentence as an authentic use of language and is not familiar with these terms, the item is context-reduced and may be more difficult than if the meanings of these terms were ignored entirely. Because of the unnecessarily difficult context, items such as these probably engage other competencies even though they are intended only to measure grammatical competence.

3.1.3 Multiple-choice tests of reading comprehension

Of the various types of multiple-choice test, the reading comprehension test has, in my opinion, the greatest potential for requiring communicative language performance. This is because it is the least restricted with respect to organization of discourse, propositional content, illocutionary force, and forms. There are basically two tasks in this test type: 1) comprehending a written text and 2) providing requested information based on the content of that text. The questions are generally of two types: incomplete statements and direct information questions. The type of information requested is usually both literal and inferential. Items in this test type may measure grammatical, cohesive, and illocutionary competence. Strategic competence, to the extent that this is involved in inference and drawing on relevant extra-textual knowledge, can also be measured, should this be desirable.

In general, while multiple-choice tests are highly restricted in terms of the type of performance required, I believe they can be used effectively to measure the receptive skills of listening and reading, and to measure the full range of competencies required in these two skills.

3.3 Oral interviews

The oral interview is probably the premier "direct" test of language proficiency. It is nearly the opposite of the multiple-choice test, in that it can require authentic language use, or communicative performance. While generally limited to the aural/oral channel, both receptive and productive modes can be measured, as can the full range of competencies involved in the skills of listening and speaking. The extent to which this test achieves its full potential, however, depends on the elicitation and rating procedures. The skillful interviewer will lead the candidate through a variety of topics, elicit a variety of illocutionary acts, and present a variety of contexts. Indeed, in a well-conducted interview, the candidate may nearly forget that it is a test.

All too frequently, however, the candidate's performance is rated solely in terms of grammar, pronunciation, vocabulary, and perhaps fluency. Such a rating scale fails to evaluate aspects of either discourse competence, such as cohesion and rhetorical (conversational) organization, or of sociolinguistic competence, such as appropriateness of register and naturalness.

Another common characteristic of rating scales that have been developed for oral interviews is the definition of the scale points, or levels, in terms of specific contexts and subject matter. A well-known example of this type of scale definition is that of the Interagency Language Roundtable (ILR) oral interview (formerly the Foreign Service Institute (FSI) oral interview). This scale has been adopted and expanded by such diverse organizations as the American Council of Teachers of Foreign Languages (ACTFL), in its "Provisional Proficiency Guidelines" and the Australian Department of Immigration and Ethnic Affiars, in its "Australian Second Language Proficiency Ratings (ASLPR)". This type of scale definition may be quite useful for specific situations such as those of the agencies of the U. S. government, or of modern language departments in U. S. colleges and universities, or of a large-scale migrant program in Australia. Because of the effect of context on communicative language use, however, context dependent definitions limit not only the use of such scales, but more importantly, their interpretation, to the specific situations for which they are designed. Thus, ratings on these different scales are of little use for comparative purposes. How comparable, for example, are the following scale definitions?

> Can handle with confidence but not with facility most social situations including introductions and casual conversations about current events, as well as work, family, and autobiographical information; can handle limited work requirements, needing help in handling any complications or difficulties.
>
> (Lowe, 1980:1-5)

> Can narrate, describe, and explain in past, present, and future time. Can communicate facts--what, who, when, where, how much--and can explain points of view in an uncomplicated fashion, but cannot conjecture or coherently support an opinion. Can talk in a general way about topics of current public interest (e.g., current events, student rules and regulations), as well as personal interest (work, leisure time activities) and can give autobiographical information. Can make factual comparisons, such as college life vs. high school life. . . . Can make a point forcefully and communicate needs and thoughts in a situation with a complication (e.g. calling a mechanic for help with a stalled car, explaining suspicious-looking possessions to a customs official)."
>
> (ACTFL, 1982:)

> Has restricted register flexibility though, where a specialist register has been experienced, will have acquired some features of it. . . . Can give detailed information about own family, living conditions, educational background; can describe and converse on everyday things in his environment (e.g., his suburb, the weather); . . . can communicate on the spot with fellow workers or immediate superior (e.g., ask questions about job, make complaints about work conditions, time off, etc.).
>
> (Ingram and Wylie, 1981:14)

The first quotation is from the definition of level two on the ILR rating scale; the second is from the ACTFL "Provisional Proficiency Guidelines" definition of the "advanced" level, which is considered comparable to level two on the ILR scale, and the third is from the ASLPR definition of level two. While the ILR statement is quite general, the ACTFL and ASLPR statements are much more specific. The problem this creates for comparability of ratings is that one is not certain that interviewers using the different scales would try to elicit exactly the same language functions (narrating, describing, explaining, explaining, conjecturing, supporting an opinion, making factual comparisons, communicating needs and thoughts). Nor is it likely that the same content areas (work, leisure time activities, student rules and regulations, college and high school life) or social registers will be of relevance to American college students, employees of U. S. government agencies, and immigrants to Australia. If there is this much potential for difference in elicitation and interpretation of ratings from scales such as these, which are very closely related in terms of their development, comparability of interpretation is even less attainable when quite different types of scales are examined.

A more generalizable and interpretable approach to scale definition, I believe, is to define levels on separate scales in terms of the characteristics of the various components of communicative competence. In the Oral Interview Test of Communicative Proficiency in English (Bachman and Palmer, 1983), for example, level two is defined in terms of several components, such as:

> Grammar
>
> > Limited range of both morphologic and syntactic structures, but with some evidence of systematic rules; control of some structures used, but with many error types
>
> Cohesion (part of discourse competence)
>
> > Moderate cohesion, including coordination and simple subordination; sometimes confusing relationships among ideas
>
> Sensitivity to Register (part of sociolinguistic competence)
>
> > Evidence of two registers and control of either formal or informal register
>
> (Bachman and Palmer, 1983:2-4)

Interviewers are instructed to elicit topics, illocutionary acts, and sociolinguistic registers appropriate to the context and to the candidate's needs and interests. Thus these factors do have an important effect on the communicative performance elicited in the interview. But since the scale definitions are independent of context and subject matter, the interviewer is not constrained to elicit a particular set of discrete grammatical structures, vocabulary items, or speech acts. This is not to claim, however, that defining such scales is not problematical. On the contrary, the identification and ranking of illocutionary acts in terms of appropriateness and level, for example, is extremely complex. Nevertheless, I belive that this approach to scale definition has a great deal of potential for providing a "common yardstick" for rating any given speech sample in terms of the components of communicative language proficiency.

3.3 Cloze tests

The cloze continues to be an enigma. While it appears to approximate quite closely the kind of processing involved in reading, and thus to involve communicative performance, it nevertheless is generally perceived by test takers as a highly artificial task. Indeed, much of the research with variations in this procedure has been motivated, in part at least, by the desire to overcome its lack of appearance of validity. I believe that this perceived artificiality is largely a function of the random deletion procedure, which frequently results in items that are nearly impossible to complete. Perhaps time and conditioning will be the ultimate solution to this problem. After all, even the multiple-choice test, which is

now widely accepted by test takers as the prototypical test, was once perceived as totally lacking in "face validity".

From my own research with the cloze I am convinced that it can measure the full range of competencies involved in reading. The key to this potential, however, lies in the specific words that are deleted, and to assure that the specific competencies one wishes to measure are in fact measured, it is essential to abandon the random deletion procedure for a rational one in which the test developer selects the words to be deleted according to criteria defined in the content specifications of the test.

4 Implications for measurement theory

Given the range of language performance required and the competencies measured by language proficiency tests, it would seem useful to consider the extent to which such tests can be adequately analyzed by current psychometric theory. One assumption of test theory, both classical true-score and latent-trait models, is that test items are locally independent (Rasch, 1960; Lord and Novick, 1968). This means that the probability of an individual's answering an item correct is a function only of his or her ability level and the difficulty level of that single item. For this assumption to be satisfied, test developers must write and arrange items so that they are as independent of each other as possible in terms of the tasks required and content included. This is clearly at odds with communicative language performance, in which the "items" of discourse are by definition related to each other and to a given context. (It should be noted that Holland (1981) has suggested a less restrictive assumption, that of "local nonnegative dependence", that may provide a practical means for determining whether item response models fit a given set of item data.)

A second assumption of currently available latent-trait models is that the test items comprise a unidimensional scale, that is, that they all measure a single trait or ability (Lord and Novick, 1968; Lord, 1980). This assumption would also appear untenable, not only in terms of current theories of language proficiency, but also in light of recent research in language testing, which indicates that language proficiency is multidimensional (e.g., Swinton and Powers, 1980; Bachman and Palmer, 1981, 1982; Dunbar, 1982; Carroll, 1983; Upshur and Homburg, 1983). As with the assumption of local independence, attempts by test developers to satisfy the assumption of unidimensionality may well result in items that are artificially restricted in their form and content. In fact, the quintessential "discrete-point" item might be regarded as unidimensional.

If current theoretical frameworks and research describe communicative language proficiency as comprising several distinct but related traits, and communicative language performance as occurring in the context of discourse, with interrelated illocutionary acts expressed in a variety of forms, it would seem that language tests would provide both a challenge and an opportunity for psychometricians to test the assumptions of current models and to develop more powerful models if necessary.

5 Implications for test development

There would appear to exist a similar challenge and opportunity for test developers to find more creative test procedures and formats. One such procedure, a variation of the dictation called the "copy-test" (Cziko & Lin, 1984), involves the visual presentation of material, and has considerable potential as measure of text processing. It is not unreasonable to expect that advances in microcomputer technology, along with its increasing availability, will provide the means for making this testing technique feasible for large-scale testing in the next few years.

Within the multiple-choice framework, I believe it would be useful to experiment further with items in which some of the distractors are partially correct. For example, the key response would be completely correct in terms of syntax, cohesion, coherence, and perhaps register, while the distractors might be syntactically correct, but not cohesive, syntactically and cohesively correct, but not in the appropriate register, and so forth. From items such as these it might be possible to derive scores for these different aspects of communicative competence. This type of item has been examined by Farhady (1980).

6 Conclusion

In this paper I have presented a framework for examining performance on language tests and have attempted to demonstrate how this framework might provide some insight into the types of language performance elicited and the language competencies measured by such tests. At this point I would like to venture some opinions regarding the extent to which language tests can or must comprise measures of communicative competence. First, I believe that it is possible for tests that do not involve communicative performance to measure some aspects of communicative competence. Second, it is quite possible that not all the traits of communicative competence are equally relevant to the language use needs of a given group. The ability to use language to perform imaginative functions, for example, is probably of less importance to college students, unfortunately, than the ability to perform ideational functions such as defining, describing or arguing. Finally, it may well be that not all the relevant abilities of communicative competence are measurable within the limitations of any given testing program.

A genuine commitment on the part of language testing researchers and language test developers to content and construct validity requires the constant re-examination of the objectives of our tests and a reassessment of the techniques we employ for eliciting communicative language performance. The innovations that result from this re-examination will have implications for both test development and test theory, in that they will require creative applications of current models and technology, and may stimulate the creation of new models and new technology.

I disagree with those who feel that the most pressing need is for further development of theoretical models. I believe that current models are sufficiently well-defined to permit empirical verification. I also disagree with those who disparage ever being able to characterize all the components of communicative competence, for such a position would lead us to cease empirical research. The highest priority, I believe, is for us to attempt to move to the paradigmatic stage of scientific development and begin empirical verification of a current theoretical model of communicative language proficiency. Furthermore, I believe that the most effective means of such empirical verification is through language testing research.

Bibliography

American Council of Teachers of Foreign Languages (ACTFL). 1982. ACTFL provisional proficiency guidelines. Hastings-on-Hudson, New York: ACTFL.

Bachman, L. F. and A. S. Palmer. 1981. "The Construct Validation of the FSI Oral Interview". Language Learning 31, 1:67-86.

- 1982. "The Construct Validation of some Components of Communicative Proficiency". TESOL Quarterly 16, 4:449-65.

- 1983. Oral interview test of communicative proficiency in English. (Manuscript).

- 1984. "Some Comments on the Terminology of Language Testing". In Rivera, C. (ed). Communicative competence approaches to language proficiency assessment: Research and applications. Clevedon, England: Multilingual Matters.

- Forthcoming. Basic concerns in language test validation. Reading, Mass.: Addison-Wesley.

Bialystok, E. and E. Ryan. Forthcoming. "A Metacognitive Framework for the Development of First and Second Language Skills". In Forrest-Pressley, D. L., G. E. MacKinnon and T. G. Waller, (eds). Meta-cognition, cognition and human performance. New York: Academic Press.

Campbell, D. T. and D. W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-multimethod Matrix". Psychological Bulletin 56, 2:81-105.

Canale, M. and M. Swain. 1980. "Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing". Applied Linguistics 1, 1:1-47.

Carroll, B. J. 1980. Testing communicative performance: An interim study. Oxford: Pergamon Press.

Carroll, J. B., 1961. "Fundamental Considerations in Testing for English Language Proficiency of Foreign Students". In Center for Applied Linguistics. Testing the English proficiency of foreign students. Washington, D. C.: Center for Applied Linguistics.

- 1983. "Psychometric Theory and Language Testing". In Oller, J. W. (ed). Issues in language testing research. Rowley, Mass.: Newbury House.

Chomsky, N. 1965. Aspects of the theory of syntax. Cambridge, Mass.: Massachusetts Institute of Technology Press.

Cziko, G. and J. Lin. 1984. "The Construction and Analysis of Short Scales of Language Proficiency: Classical Psychometric, Latent-trait, and Nonparametric Approaches". TESOL Quarterly 18, 4:627-47.

Dunbar, S. B. 1982. "Construct Validity and the Internal Structure of a Foreign Language Test for Several Native Language Groups". Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Faerch, C. and G. Kasper. 1983. "Plans and Strategies in Foreign Language Communication". In Faerch. C. and G. Kasper (eds.) Strategies in interlanguage communication. London: Longman.

Farhady, H. 1980. Justification, development, and validation of functional language tests. Unpublished Ph.D. dissertation, University of California at Los Angeles.

Halliday, M. A. K. 1976. "The Form of a Functional Grammar". In Kress, G. (ed). Halliday: System and function in language. Oxford: Oxford University Press.

Holland, P. W. 1981. "When Are Item Response Models Consistent with Observed Data?" Psychometrika 46, 1:79-92.

Hymes, D. H. 1972. Towards communicative competence. Philadelphia: University of Pennsylvania Press.

Ingram, D. E. and E. Wylie. 1981. Australian second language proficiency ratings (ASLPR). Brisbane: Australian Department of Immigration and Ethnic Affairs.

Lado, R. 1961. Language testing: The construction and use of foreign language tests. London: Longman, Green & Company.

Lord, F. M. 1980. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. and M. R. Novick. 1968. Statistical theories of mental test Scores. Reading, Mass.: Addison-Wesley.

Lowe, P. 1980. manual for LS oral interview workshops. Washington, D.C.: Language School, CIA.

Munby, J. 1978. Communicative syllabus design: A sociolinguistic model for defining the content of purpose-specific language programmes. Cambridge: Cambridge University Press.

Rasch, G. 1960. Probabilistic models for some intelligence and attainment tests. Copenhagen: The Danish Institute for Educational Research.

Savignon, S. J. 1983. Communicative Competence: Theory and classroom practice. Reading, Mass.: Addison-Wesley.

Swinton, S. S. and D. E. Powers. 1980. Factor analysis of the test of English as a foreign language for several language groups. TOEFL Research Reports, Report 6. Princeton, N. J.: Educational Testing Service.

Upshur, J. A. and T. J. Homburg. 1983. "Some Relations among Language Tests at Successive Ability Levels". In Oller, J. W. (ed). Issues in language testing research. Rowley, Mass.: Newbury House.

Widdowson, H. G. 1978. Teaching language as communication. Oxford: Oxford University Press.

MICROCOMPUTERS AND LANGUAGE TESTING

Terry Culhane (University of Essex)

Computing and Language Teaching

The art of program writing and the art of language teaching came together many years ago with PLATO and such devices, which were dependent on large-scale "main-frame" machines[1]. Smaller computers, such as, for example, the one at the University of Surrey, were devoted entirely to the teaching process, and programs such as those devised by G.Corbett,K. Ahmad and others for the teaching of Russian numerals[2]. The microcomputer revolution, with its rapidly increasing versatility, availability and accessibility, inevitably increased the range of programs available. Some of these, such as the Carbondale Russian Irregular Verbs and French Irregular Verbs are encyclopoedic, comprehensive and rather traditional in their approach. They are, nevertheless, very effective in the classroom. Students seem to like the stimulus brought on by the interaction between themselves and the TV screen. It is inevitable that many of the early programs used a "programmed learning type of approach, and that many of the testing devices used were of the multiple-choice type. Cloze tests, gap-filling exercises and "storyboard" are also favourite exercises, and some software firms, such as WIDA[3], provide so-called "authoring" packages, which enable the teacher to write his own material. It is likely that these popular and effective techniques will be complemented by programs devoted to "information gathering" and language games, such as the highly imaginative "Jeu de ménages" and "Quelle tête" of Barry Jones [4]. Many of the programs being produced for the latest BBC French series have features not previously seen in language teaching software. These will be available later this year and contain about 45 language teaching programs by Mike and Marie-Madeleine Kenning[5].

There are, of course, many ways in which a computer can be used in the classroom. Many teachers have used the computer as a source of information which the student can refer to when appropriate. This technique is popular in primary schools, where the children have to gather information from several different sources, and is often used with children working in small groups. A second method is to use the computer as a source of stimulus for the whole class. "Storyboard" and "Crossword" are particularly good for this purpose. The most potent use of the micro-

computer is likely to be for self-instructional purposes, or, at least, using self-instructional techniques in small groups.

Self-instructional techniques and testing

Many commercially available programs contain materials of a "testing" nature. The WIDA vocabulary programs, for example contain multiple choice tests of vocabulary. These would not pass scrutiny as tests, since their distractors are drawn at random from sets of words at different levels. Other self-instructional materials[6] make extensive use of "testing" as a teaching device. The validity of such tests should be estimated on the basis of their success as teaching materials, and not using criteria intended for the validation of tests proper. Self-assessment is there in order to satisfy the person being assessed, who is both examiner and examinee, and if devices are didactically effective they may be said to have met the criteria for which they were devised. In other words they have the only kind of validity they need. Courses which rely on information gathering techniques based on comprehension can claim high "authenticity", incidentally, and if teaching and testing go hand in hand, the branch of teaching known as self-insruction is the one closest to testing. It is here that testing techniques are given maximum effect as teaching devices.

BASIC and the Teacher

Most microcomputers operate using a language called BASIC, and all of the programs quoted above are written in that language. We at Essex run two courses in CALT for our MA students in Applied Linguistics, and this year we have successfully run a course in the Certificate of Applied Linguistics and the teaching of English (CALTE). We have discovered through (sometimes painful) experience, that not all students are capable of writing computer programs of great sophistication. All of them succeed to some extent, but the amount of time needed to write programs of publishable standard is greater than the average teacher is prepared to spend. I do not wish to discourage anyone from attempting to write programs in BASIC. Quite the reverse! The least that you will learn from the process is that the medium has its virtues and its limitations, and that different computers have their own variations of BASIC "built-in" to their operating systems. There are, however, easier ways of learning to program than by learning BASIC. One is through an "author" language specially devised to act as an interface between the teacher and the computer. There are several such languages, and we have one of them at Essex. It is used on two of the courses mentioned above.

EXTOL

The language EXTOL[7] was invented by Dr. M. Kenning of the University of Essex and his wife, who teaches at the University of East Anglia. Within four hours the more able students are writing teaching programs. The language uses a "driver" program written in PASCAL or FORTRAN and is purpose-built for language teaching. It is also a very powerful testing device, and I would like to show you how it works and how we intend to use it.

EXTOL, like BASIC, works through a program line by line from the top. It looks at the first character in each line and the nature of that character determines what the computer does. There are four types of line, and therefore four characters to distinguish them. I shall describe them by writing a program for you. The marker % indicates a comment line which is ignored by the computer and which is there to help the program writer remember what he has done. Here we go:-

```
%Remember that this line is ignored. I am now going
%to tell the computer to put something on the screen.
%I do this by leaving a letter in the first space.
What is the capital city of Finland?
%The above is a "text" line. We now want to accept
%whatever the student puts in. We do this with a
%"command" line, which starts with a dot.
.accept
%Now you want to compare his answer with the correct
%one. You do this with another command.
.match(1,1).Helsinki.
%This means if he gets it right put one in counter 1.
%Then we send the computer to counter 1 (there are
%nine such counters. If it discovers one in there it
%knows that the student answered "Helsinki". If there
%is nothing in there, he did not. We do all of this
%with one instruction-
.goto(1)..good.
%This means if he got it right jump to a "label" line
%called good and go on from there. A label line is
%marked with a colon in the left-hand column. If he
%gets it wrong he falls through to the next line.
I'm sorry, but you got it wrong.
%If you wanted to, you could send him back to the
%beginning by another "goto" line and another label
%line. Or you could stop him, like this:-
.stop
%He gets to the following line if he got it right:-
:good
Well done! Ask the teacher for your prize!
.stop
```

EXTOL has seventeen commands only. It is possible to

score a student's performance in a test by using the command .increment(5,1). This would increase the value of counter 5 by one. Any one of the nine counters may be used and the teacher has the choice of whether or not to inform the student of his score. Student responses are also recorded in a "response" file, which is not accessible to the student. A full list of EXTOL commands is in Appendix A. We have constructed multiple-choice, gap-filling, cloze tests and even C-tests using EXTOL. Here is an example of how EXTOL can be used in a multiple-choice test.

```
.response lulu
.reset(1)
.reset(2)
.define.multi.
<Q>
<1>
<2>
<3>
<4>
.accept
.match(1,1).<ANS>
.goto(1)..inc.
.exit
:inc
.increment(2,1)
.end
.perform.multi.
Q=He used to go fishing once a week, but he ... now
1=A is almost never going
2=B is never going
3=C never goes
4=D never is going
ANS=C
.perform.multi.
Q=Never....him looking so ill.
1=saw I
2=I have seen
3=have I seen
4=I saw
ANS=c
Well done, you got
.print(2)
right
.stop
```

The above program will present a question as often as the command .perform.mult. is repeated and new parameters (questions and answers) are fed into the computer. The student's responses are stored in a file called "lulu", and in this variant the student is told his score. This example was written for purposes of illustration. We also have a series of tests written in EXTOL which are more

sophisticated and which make much better use of the medium. Most of these are variations on the cloze technique and we hope to report on them at some future conference.

Other Ways in which Computers might Help Testing

It would be a wasted opportunity simply to feed into computers multiple-choice tests, cloze tests and the like, and use the machines as vehicles for existing methods. One of the virtues of computers is that they give us the opportunity of presenting our test material in a new way, by giving the student as much as we want for the length of time we want him to be exposed to it. This in itself should give rise to new ways of testing. The way in which we process the students' responses might also be different. Computers will enable us to test aspects of language we have so far not been able to test. The experiment at Essex mentioned above is attempting to find out something about the relationship of lexical availability and performance in cloze tests. Other factors which come to mind include speed of response, speed of understanding, availability of structures for productive purposes. All of these might well be investigated using microcomputers. Teachers and testers should be asking themselves the question, "What can I test with a computer that I cannot test by other means?" It is possible that the answer to this question could provide the most exciting advances in knowledge about how we learn languages that we have had in our lifetime.

APPENDIX A

Commands in EXTOL

.ACCEPT	Accept input from keyboard
.DEFINE.drill.	First line of exercise called "drill"
.END	Last line of above exercise
.EXACT	Set matching to accept as correct exact response from student
.EXIT	EXIT from drill or exercise
.FEED(5)	Type 5 (or more if you wish) blank lines to screen
.GOTO(1)..XXX.	If the value of counter 1 (there are 9 such counters) is zero carry on to the next line. If it is not zero go to a label called XXX
.HELP filename	Specify name of "help" file
.INCREMENT(1,1)	Add 1 to value of counter 1
.KEYWORD	Set matching to keyword mode
.MARK	Save current value of score counters in response file
.MATCH(1,1).ans1.ans2.	If answer 1 or answer 2 occurs as student input add 1 to the value of counter 1
.PERFORM.dril.	Perform an exercise (or test, etc.) called "dril"

.PRINT(1)	Print the value of counter 1 on the screen
.RESET(1)	Set the value of counter 1 to 0
.RESPONSE filename	Specify the name of the response file
.STOP	Stop the lesson

Bibliographical notes

1. PLATO (Programmed Logic for Automatic Teaching Operation) Control Data Corporation.

2. Ahmad,K., G.G. Corbett and J.M. Edwards.1978. "Developments in computer assisted learning for Russian". Russian Language Journal 32:37-42.

3. WIDA Software, 2 Nicholas Gardens, London W5 5HY

4. Jones, Barry, 1984 "Quelle tête" and "Jeu de ménages" two programs available from Cambridge University Press. Similar programs are also available in German.

5. Kenning,M.J.,andKenning,M-M.,"A vous la France" A series of programs to accompany a BBC French television series. Available mid-1985.

6. Culhane,P.T. "Russian Language and People" BBC Publications, 1980.

Culhane & Chartier,J-P.1984. "English Alone"Pergamon Press and Omnivox (Paris).

Culhane,P.T.& Welsh,M. 1985 "The Russian Aphabet".In preparation.

7. Kenning,M.J.,& Kenning, M.M.,1982"EXTOL: an approach to computer assisted language teaching"Association for Literary and Linguistic Computing Bulletin"10:8-18.

See also, by the same authors, a guide to writing language teaching in BASIC- "Introduction to Computer Assisted Language Teaching"Oxford,1983.

EVALUATION OF THE PRODUCTIVE SKILLS IN THE SECOND LANGUAGE: THE WORK OF THE ASSESSMENT OF PERFORMANCE UNIT (APU)

Peter Dickson (National Foundation for Educational Research)

1 Background

In carrying out the large scale testing necessary in a national survey -and the first APU survey of foreign languages involved about 12,000 pupils in schools in England, Wales and Northern Ireland - many different factors have to be reconciled. The approach to the problems of test construction, assessment and interpretation needed to be pragmatic and not theoretical. The framework devised for the assessment does draw, of course, on language testing and language learning theory, but what I propose to describe in outline is the development of a rationale for testing, the elicitation procedures used in the assessment of speaking and writing, the methods of assessing performance, and the application of the results. There is not sufficient time here to give an account of the results being obtained. I shall restrict myself, therefore, to those aspects of the results which throw light on the effectiveness of the methods adopted.

2 Monitoring at age 13

I would like to say a word or two at this point about the background of the APU which was set up in 1974 by the Department of Education and Science in England and Wales. In the mid 1970s work began on the national monitoring of certain areas of the curriculum. Mathematics was monitored first, followed by English and then Science. The aim of the monitoring at the time was to obtain detailed information about pupil achievement in a situation in which individual schools made independent decisions about the content of the curriculum. The monitoring of foreign languages began in 1981. It was decided to monitor French, German and Spanish as a first foreign language in the first of the cycle of surveys in 1983 and to monitor French only in the subsequent surveys scheduled for 1984 and 1985. Another decision made at the time was to monitor pupils at age 13 only. It would not have been possible it was argued to obtain a representative sample nationally if the age chosen for monitoring had been greater, as many pupils are known to abandon studies of a foreign language at that stage. Equally, it was not thought profitable to monitor before that age as many pupils would have completed only two years' study by the age of 13.

At the beginning of the project the monitoring team had to consider the situation in schools. We also had to take account of the pattern of learning among pupils in the relevant age group. Teachers in schools had begun to seek new approaches to deal with the problems presented by the comparatively new situation in which pupils of all attainments were now learning a foreign language. This was one of the problems we had to confront in developing suitable instruments; the variety of teaching approaches used was another,and these problems were found to be especially acute when we came to assess language production. Opportunities for pupils to develop speaking skills in school were clearly limited. It was also claimed by many teachers that no writing in the foreign language

was done in the early stages, and some insisted that it was inappropriate to assess it.

3 Rationale for testing

Before the monitoring team was appointed, a working group set up by the APU had developed and published guidelines for the testing programme and the monitoring team took these as its first reference point. The predilection of the monitoring team for language tests with a communicative orientation was consistent with the proposals made by the working group. The guidelines set out a scheme thought relevant and this defined the linguistic content and sociolinguistic context of the kinds of test which they proposed. This meant that areas of the foreign language were specified in terms of topics, functions and general notions, and that the situations for the use of language were indicated in conjunction with the language skills required. No tasks were specified in the guidelines.

The priority, as we saw it, was to describe the testing domain by drawing up a range of settings and a list of tasks considered appropriate for the development of the tests. This had the obvious advantage of providing a rationale for identifying sources, both for texts to be used in the listening and reading tests, and for specific speaking and writing activities. This was a rationale which served also the interests of efficiency and economy, as it ensured that meanings focused on by the questions in the listening and reading tests were those which were implicit in the tasks, and that the speaking and writing activities were those which were most relevant in the chosen settings.

We have heard support in the course of this conference for testing discrete aspects of communicative performance. This explains our compromise position perfectly. To begin with we had no alternative but to accept the traditional four skill model of testing. Other approaches could not satisfy the rigorous requirements of measurement. Besides, the limited experience of language learning which characterised the population of 13 year old pupils in schools means that appropriate integrated tasks could not be identified in sufficient numbers for our purposes.

4 Framework for the writing and speaking tests

In order to carry out the monitoring, five tests in each of the areas of writing and speaking were developed. The five tests in each area are designed to cover a range of topics, and any one test is designed to cover a range of skills. Another feature of the construction which is common to both the writing and the speaking tests is the sequencing of tasks. At the beginning of each test, the tasks which pupils are required to carry out are closely structured. Subsequent tasks are progressively more open-ended. Some may argue that the final tasks still do not offer scope for a fully independent performance, but they should bear in mind the character of the target population.

In making decisions about the content and organisation of the tests, there were three principal considerations. First, we wished to satisfy ourselves that the tasks selected for inclusion in the tests were both useful and appropriate. That is to say that they were tasks which were accessible to the pupils both in terms of their personal experience as well as in terms of their language learning experience. Secondly, we needed to be sure that the tests were valid in terms of content. The skills required to carry out the tasks, and the topics, needed to be those which commonly formed part of the curriculum for 13 year old pupils. The third was a particular concern of the monitoring. The tests needed to be organised in such a way as to maximise the pupils' opportunities to demonstrate the range of skills which they possessed. Put another way, the tests needed to yield evidence of positive achievement at the lowest level, and yet impose no ceiling on levels of performance which might be found among 13 year olds. While performing this balancing act, we had also to make as much of each test as possible accessible to the lowest achieving pupils who should not, we thought, be alienated by the experience of taking part in the survey.

I should like to summarise the decisions taken to meet these concerns by describing the framework and content of the writing and speaking tests.

In the writing tests the tasks selected are centred on aspects of letter writing. Although other kinds of tasks such as message writing were considered, it was decided that letter writing was particularly suitable because it is an activity which is familiar to pupils in the age group, and also because it offers the flexibility necessary to introduce into the writing tests a range of topics.

Decisions as to which topics to include were less easy to arrive at, and were only settled after extensive rounds of pre-testing and a large scale pilot survey. In the final form of the tests there is a common component which is based on the topic of personal information, and another section specific to each test,which contains tasks based on one of a range of topics judged to be those with which 13 year old pupils are most familiar. Each writing test is in fact based on a single landscape picture which focuses on the topic, providing unity to the test as a whole.

How were we to maximise opportunities for pupils to write? How were we to avoid imposing an artificial ceiling on performance? One solution has already been referred to - the inclusion of both closely structured and open-ended tasks. Another solution was to ensure that pupils needed to draw on a reasonably wide range of skills in order to perform the different tasks included in the tests. Each writing test requires pupils to perform the skills of copying, listing, giving personal information, describing and reporting. The first of these, copying and listing, reflect activities which pupils commonly engage in in class and were included in the writing tests to ensure that the least able pupils have at least some tasks which are accessible to them. It is recognised that the decision to include these activities partly compromised the rationale which dictated that only realistic tasks or real life tasks should be included in the tests. It is relevant to note

here also that in order to elicit even the most modest attempts at writing, the tasks are presented in the test booklets in such a way as to encourage pupils to write something, even if what they write is incomplete or if they know they are making mistakes in the process of writing. It is not easy, of course, to know the extent to which that particular aim has been realised.

I turn now to the framework and content of the speaking tests. The tasks selected for inclusion in the speaking tests are drawn, as for other areas of the testing, from what were considered appropriate settings - travelling in a foreign country as a family, or in a school group, entertaining or visiting a person of a similar age from a foreign country. Within these settings, the tasks involve taking part in conversations, both social and transactional, and giving a more extended description or account. In order to accommodate this range of speaking tasks, each test includes a section requiring the pupils to read aloud, a closely structured conversation, two role tasks and a task requiring more extended speaking - a description or report of events. The point has already been made in respect of both the writing and the speaking tests that the organisation of the tests reflects progression from the closely structured to the open-ended. This is especially marked in the speaking tests in which the degree of interaction possible increases through the test, as also does the degree of independence required of the pupil.

As in the writing tests the content was settled as a result of extensive pre-testing and piloting of the material. In their final form the speaking tests have a common component, again on the pattern of the writing tests. This is the structured conversation which, in the interests of realism, or to mitigate its artificiality, has its own natural development, and a role scripted in French which needs to be played by the assessor. More relevant here is the content of the structured conversation which is designed to elicit many of the utterances which it is thought pupils might need to produce in initial encounters with foreign speakers. The linguistic content of the utterances cued by the role tasks is also to some degree prescribed. However, these are tasks in which a degree of interaction is possible. The inclusion of two role tasks in each of the speaking tests also ensures that there is a balance between the requirement to respond and to ask questions. The final task in each speaking test requires pupils to give a more extended account. Here the opportunity for pupils to demonstrate communicative strategies is supposedly at its greatest. Results, however, have shown lack of independence on the part of most pupils performing these tasks, compensated only by the answers to questions which the assessors are briefed to put to assist them.

I have given these details of the content of the writing and speaking tests, as illustration of the methods by itself would not convey an impression of the overall framework. I hope it will also provide a perspective for the description I shall give of the assessment criteria. It should be clear by now that one of our initial concerns in developing the tests was to satisfy the concerns of face validity. In devising the methods of assessment, we needed to take account not only of the rationale of the test construction, which placed tasks at the centre of the development, but also of the need to provide as much detailed description as was possible of pupils' performance.

5 Assessment criteria

The procedures devised for assessing performance of writing and speaking had to fulfil more than one function. They needed, on the one hand to provide overall scores, and on the other evidence of positive achievement at different levels. The solution was to develop scales of performance which provided not only grades for different levels of performance but also descriptions at each of the levels.

Below is an example of one scale used to assess the production of a written report of events.

5 Performs the task fully. Uses a range of appropriate structures and vocabulary. Generally accurate.

4 Performs the task adequately. Generally uses a range of appropriate structures and vocabulary. Inaccuracies do not obscure meaning.

3 Performs the task at a minimum level. Uses a limited range of structures and vocabulary, sometimes inappropriately. Inaccuracies may obscure some meanings.

2 Does not perform the task adequately. Uses a very limited range of structures and vocabulary, often inappropriately. Generally inaccurate.

1 Writes only a few words relevant to the task. May resort to English.

0 Nothing relevant to the task set.

- No attempt.

The descriptions were produced by reference to the broad criteria of communicative effectiveness, appropriateness and range. Effectiveness was considered to be the extent to which a pupil completed the task. Appropriateness included grammatical accuracy and coherence - phonetic accuracy and fluency in the case of speaking. Range related to the range of vocabulary and structure appropriate to the task. It soon became obvious in the development that a single scale of performance would not be applicable to the whole range of tasks. The differing nature of the tasks meant that the emphasis placed on any one criterion would not always be the same. For example, the parameters of the closely structured tasks were clearly set out by the separate cues, and range (or quantity) was inevitably a more crucial factor here in an effective performance.

The six points on the scale were arrived at by referring to the evidence of the pilot survey. This confirmed our view that six was a natural number of steps to include on the scale, and consideration of norms at that stage helped us to differentiate one from the other. Although the scales of performance provided descriptive categories, their main function was to facilitate rapid and reliable scoring. It is for this reason that the categories are economically expressed. It was found in the early stages of development that attempts to characterise

performance in detail by reference to a greater range of criteria tended to exclude large numbers of performances from classification. The simple description seemed in practice to be much more easily applicable. To summarise - the scales of performance are not arrived at by external standards but by reference to the variety of writing and speaking produced by the population surveyed. They are task-related. They provide broad descriptions of performance at each level. Most important, they provide a clear focus for distinguishing levels of performance. I shall return to the application of the scales for the purpose of scoring writing and speaking performance, but first I should like to complete my account of the procedures devised for carrying out the evaluation.

6 Detailed Analysis

In an attempt to elaborate on the grade descriptions, a system of detailed analysis of pupils' writing and speaking was worked out. Briefly - in the first survey the emphasis of the detailed analysis was on the components of the task performed, on the forms that were used to perform them and on the kinds of errors that were made. This kind of information was obtained on a selection of tasks, that is to say on one each of each type of task. The information was gathered in the following way. A list of items was entered on a grid by reference to a sample of the pupils' writing and speaking and the incidence of the use of each of these was entered by hand. The information obtained in this way provided the basis for producing detailed descriptions of performance of specific tasks. It was less easy however to use the information to make statements about different levels of performance generally. The approach to the detailed analysis will be further developed when the results of the 1984 survey are processed. Schemes which we are applying at present focus on other aspects of extended text creation in both writing and speaking. In writing we are aiming to obtain precise information about length, range and complexity, for example, in relation to levels of performance, and in speaking the schemes are directed in some cases towards the communicative strategies which pupils use in extended speaking, and in other cases, at the degrees of effectiveness and appropriateness pupils demonstrate in response to specific cues. This may help us to examine, in a systematic way, the difficulty pupils may have with different types of speaking and writing tasks. The national monitoring is of course a diverse exercise and, although we are collecting a rich set of data, we are conscious of the fact that it is not possible at present to do more than explore the possibilities of it in a general way.

7 Interpretation of Results

I would like to just summarise the uses to which the information obtained by the procedures I have described were put. First the unweighted scores obtained by placing performances on the scales I have described were aggregated to produce overall performance figures. These of course were crucial for reporting on performance in different modes of the assessment. Secondly, we were able to look at the distribution of scores across the scales of the performance both on individual tasks and on types of task for different groups of pupils. By reference to the evidence produced by the detailed analysis, we were able to elaborate descriptions of performance of specific tasks at different levels. Finally, we were able to review the evidence of the detailed analysis across the whole range of tasks in relation to the pattern of

distributions in order to produce generalised descriptions of peformance at least at high, average and low levels. The starting point for these generalised statements were the scales of performance which were related to each other at the same level, even if they were articulated differently to take account of the different tasks.

8 Reliability

It is perhaps not appropriate to complete this account without referring to the measures we took to control the sources of variation present in any assessment of writing and speaking. It will be clear to you from what I have said that in developing the assessment materials a major concern was to demonstrate high content validity, and face validity. However, when we had completed the development of the tests with these concerns in mind, it became of crucial importance to investigate the question of reliability. In the course of pre-testing, and in the period between the pilot survey and the main survey, a number of measures were taken to increase the reliability of the tests. These improvements were brought about by reorganising the form of the tests and revising specific items, rewording instructions and so on. Results finally showed the rewards which are available if tests are painstakingly developed. In the pilot survey the reliability of some tests calculated by the Cronbach-alpha formula ranged from .3 to .9. In the main survey the reliability figures across all the tests, five in each of the areas of writing and speaking for French, German and Spanish, was in the narrower range of .85 to .9 and some were higher than that. We took this as evidence of a satisfactory degree of internal reliability. Indeed, it is hard to see how it could be improved without compromising the validity of the tests.

An equally pressing question of course is how to control for marker and inter-marker effects. The marking of speaking tests, in which the conduct of assessors is an additional variable, as we all know, presents the greatest problems. We are convinced that, following the proper development of the assessment materials, the most important task is the training of assessors and markers. The training carried out for the assessors of speaking needs also to be a feature of the training of the markers. They need to be completely familiar with the rationale and the content of the test material, and they need to understand the procedures laid down for the conduct of the tests. It is only at this stage that markers can benefit from practice in applying the scales of performance. Theyneed to understand the concepts underlying the style of assessment represented by the scales of performance, and they need ample illustration of pupils'writing and speaking to exemplify performance at different levels. Practice in applying the scales should include discussions of differences. Our own experience is that this process is valuable in standardising the approach of assessors and in producing a high degree of consistency among teams of markers. Analyses of variance continue to bear out this view, even if random fluctuations persist in the marking of speaking.

It is of interest perhaps to mention one finding which emerged from these analyses. In the 1983 survey speaking assessors were required to score pupils' performance on-the-spot. It was found that the mean of these scores was lower than the mean score of markers working from tapes for every part of the speaking test. However, in spite of the fact that

assessors appear to mark with more severity their scores correlated highly with those of markers. Pearson correlations ranged from .78 to .81 for the four parts of the speaking test.

9 Conclusion

At the outset of this project we were conscious of the conflict which existed between the need to construct writing and speaking tests based on realistic tasks relevant to the current concerns in the curriculum, and of the need to satisfy the rigorous requirements of educational measurement. We were aware if you like of the need to effect a reconciliation between the concerns of validity and reliability in the language tests we were to use in the national monitoring. In this sense we hope at least to have shown that progress is possible by adopting a process of cautious compromise, realising what is possible of the innovative while safeguarding what is essential if judgements based on our assessment are to be taken seriously.

References

Department of Education and Science (1980)
Foreign Language. Assessment of Performance Unit.

Dickson, P., Boyce, C., Lee, B., Portal, M., Smith, M. (1985 in press)
Foreign Language Performance in Schools: Survey of French, German and Spanish. Report No. 1. London: HMSO.

Appendix 1. Examples of tasks used in APU speaking / writing tests, with some scales used to assess performance.

1. Writing tests
PART A. Giving personal information.

In a letter to your French penfriend you want to write about yourself and your family.
Use the list below to give as much information as possible.
Write complete sentences if you can.
If you can only write short phrases, do that instead.

Your penfriend will want to know:
What your name is.
How old you are.
If you live in the town or country.
If you have brothers or sisters.
What job your mother or father does.
What you like doing.
What you look like.
If you will write often.

Example: What you like doing. J'aime le football.
Write the information, in French, in the space below.

_____

_____

_____

....

PART B.
1. Listing

Look at the picture.
Below are two words for things in the picture.
Copy out the words into the spaces provided.

	Copy the words here
vaisselle	_____
parapluie	_____

Look at the picture again.
Now fill as many of the spaces below as you can, using words for things in the picture.
Write ONE word in each space.
Do not use words given above.

_____ _____

_____ _____

2. Describing
In a letter to your penfrind you want to describe the inside of your house.
Use the picture to help you decide what to write about.
Write your description using comlete sentences if you can.
If you can only write short phrases, do that instead.
Write, in French, in the space below.

_____

_____

_____

....

3. Reporting
Last Saturday you decided to stay at home.
In a letter to your penfrind you want to say what you did during the day.
Write this, in French, in the space below.

_____

_____

_____

_____

...

36

PART B: Assessment scales

2. Describing

5 Describes several of the features/activities relevant to the task. Uses a range of appropriate structures and vocabulary. Generally accurate.

4 Describes some of the features/activities relevant to the task. Uses a range of appropriate structures and vocabulary. Many minor inaccuracies.

3 Describes some of the features/activities relevant to the task. Uses a limited range of structures and vocabulary. Inaccuracies may obscure precise meanings.

2 Refers to very few relevant features/activites using a very limited range of vocabulary. Structure recognisable. Generally inaccurate.

1 Writes only a few words relevant to the task. May resort to English.

0 Writes nothing relevant to the task.

- No attempt.

3. Reporting

5 Performs the task fully. Uses a range of appropriate structures and vocabulary. Generally accurate.

4 Performs the task adequately. Generally uses a range of appropriate structures and vocabulary. Inaccuracies do not obscure meaning.

3 Performs the task at a minimum level. Uses a limited range of structures and vocabulary, sometimes inappropriately. Inaccuracies may obscure some meanings.

2 Does not perform the task adequately. Uses a very limited range of structures and vocabulary, often inappropriately. Generally inaccurate.

1 Writes only a few words relevant to the task. May resort to English.

0 Nothing relevant to the task set.

- No attempt.

2. Speaking tests

PART A.

1. You are in a post office in France and you want to send this telegram to your penfriend. Read it aloud.

Destinataire	-	Vincent 9 rue de la Tannerie Poitiers
Toujours en vacances avec parents.		
Arrive lundi 13 heures gare SNCF.		

2. You are on a beach in France with your family. You want to find out from a French girl where you can buy an ice-cream. The assessor will play the part of the French girl.

1. Say hello.
2. Ask her if there is a café.
   She says there is, and she is going there too.
3. Ask her where the café is.
   She says it's at the end of the beach. You walk along together.
4. Ask her if she spaks English.
   She prefers to speak French.

:::::::

12. Say thank you very much.
    She is going to leave the beach.
13. Say good-bye.
    She asks if you come to the beach every day.
14. Say yes, you will be there tomorrow at 10 o'clock.

PART A/2: Assessment scale

Structured conversation

5 Can convey the message in all cases, using expressions which a native speaker might use in the context.

4. Can convey the message in all or most cases, almost always using appropriate expressions.

3. Can convey the message in at least half of the cases, sometimes using appropriate expressions, at other times producing incomplete utterances.

2. Can convey the message in some cases, mostly in incomplete utterances.
   or:
   Can convey the message in few cases using appropriate expressions.

1. Can convey the message in a very few cases, and generally in incomplete utterances.

0. Nothing recognisable as relevant to the task.

- Not attempted.

PART B.
In this section of the test you take part in 2 short conversations. Read the instructions for each conversation and look carefully at the pictures. These will show you what you have to say.

1. Meeting a French boy

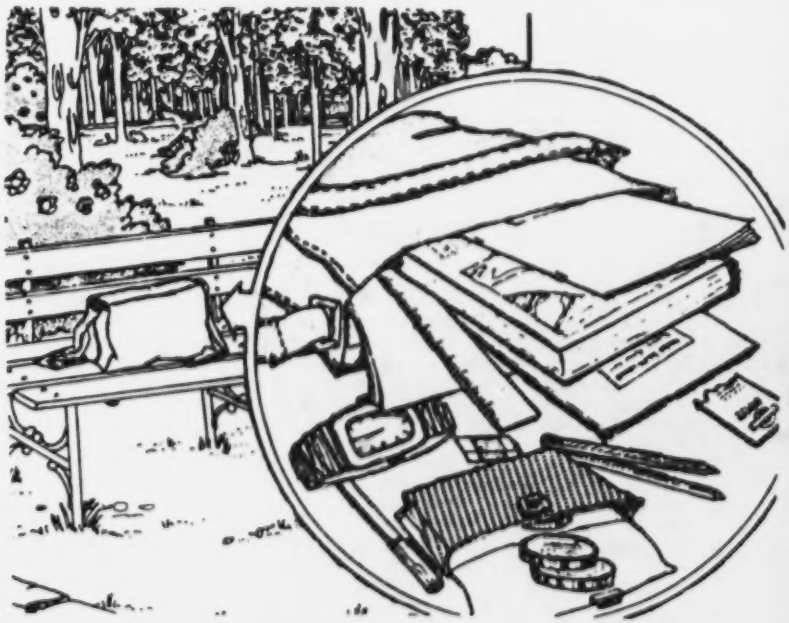
You meet a French boy at school who is on an exchange visit. You want to get to know him but he does not speak English. Find out his name and age and ask him about the different members of his family shown in the photograf h below.

The assessor will play the part of the French boy.

2. Asking about lost property

You are on a school exchange in France and you find you have left your bag in the park. You go to the Lost Property Office to ask if it has been found.

The assessor will play the part of the person in the Lost Property Office.



PART B: Assessment scale

Rôle tasks

5 Conveys or elicits all details relevant to the task. In general, grammatically accurate.
4 Conveys or elicits all or most details relevant to the task. Uses some inappropriate forms. Some grammatical inaccuracy.
3 Conveys or elicits only details relevant to the task. Uses some inappropriate forms. In general grammatically inaccurate.
2 Produces language which succeeds in conveying or eliciting few details relevant to the task.
1 Says one or two words relevant to the task.
0 Says nothing relevant to the task.
- Not attempted.

PART C.

In this part of the test you have to give a more detailed description or account.

Make sure you have the sheet of colour photographs in front of you.

You want to describe to your penfriend a place where you have been on holiday. The photographs on the sheet are of the town, its surroundings and of the activities available. Use them to describe the place, what you can do there and where you can stay.

The assessor will play the part of your penfriend.

PART C: Assessment scale

Extended speaking

5 Uses a range of structures and vocabulary relevant to the performance of the task. In general grammatically accurate. Uses appropriate tense or tense indicators.
4 Performs the task adequately, using only one or two structures but an appropriate range of vocabulary. Makes some grammatical errors which do not obscure meaning. Some attempts to use tenses (or tense indicators) other than the present, if appropriate.
3 Performs the task at a minimum level. Structures and vocabulary are limited in range, and sometimes used inappropriately.
2 Does not perform the task appropriately. Produces some relevant language, but in fragmented form. Frequent grammatical errors. May resort to English.
1 Produces only one or two relevant words or phrases.
0 Produces nothing relevant to the task set.
- Not attempted.

DEVELOPING CLASSROOM TEACHERS' ASSESSMENT SKILLS

Mary C. Kilborn (University of Stirling)

ABSTRACT

This paper contains a preliminary presentation of the work of a research project entitled "The School-Based Assessment of Communicative Foreign-Language Skills in Secondary Classes 3 and 4". The project, which is funded by the Scottish Education Department is currently being carried out in the Department of Education at the University of Stirling. The models of communicative competence and performance being developed for use with classroom teachers are presented. The general pattern of testing observed during preliminary visits to six cooperating schools is described and also the views currently held by teachers concerning the teaching and assessment of communicative skills. The future action research strategy of the project is also outlined.

The other members of the research team are R.F. Mitchell, R.M. Johnstone and D. Kent. I would like to thank them for their help and support in the preparation of this paper.

INTRODUCTION: SCOTTISH REFORMS AT 16+

The question of the role of the classroom teacher as assessor is one which has been receiving increased attention in Scotland in the past few years as a result of reforms which are taking place in the Scottisch system of assessment and certification for 16-year-olds. In 1977 the Munn and Dunning Committees reported on curriculum and assessment respectively in the third and fourth years of secondary education in Scotland. Since then have followed feasibility studies, government implementation proposals and pilot studies, with the result that the reform is well under way. New assessment and certification procedures in English, Maths and Science will be introduced in 1986; certification in French is planned for 1987 and in German and Spanish for 1988. This means that the present secondary class 2 (S2) who are learning French will be taking the new "Standard Grade" examination as will those pupils in S1 who are learning German or Spanish (although there are few of them as these languages are usually started as a second language in S3).

The Munn/Dunning programme has several significant features. The first is that the "Standard Grade" examinations will provide syllabuses and certification for all pupils at the age of 16, whereas the present "Ordinary Grade" (O Grade) examinations are intended for the more "academic" pupils only. This is being done through the drafting of syllabuses at three levels (Foundation, General and Credit). Secondly, the introduction of limited performance descriptions (called "Grade Related Criteria")

alongside the award of numerical grades represents a partial shift from norm-referenced towards criterion-referenced assessment. The third main feature, and this is the one that concerns us in this paper, is that classroom teachers will be required to take on significantly increased responsibilities in the summative assessment and certification of pupils' FL performance, which has historically been the exclusive role of the Scottish Examination Board.

Joint Working Parties have been set up in the various subject areas to make implementation proposals based on the government recommendations. The Modern Languages Joint Working Party, which first met in October 1982, published its report in June of this year. The proposals contained therein represent a significant response to criticism that the present "O" Grade FL examination shows a one-sided concern with learnes' mastery of the formal target language system, and, in particular, that it neglects oral communication skills. The proposed new scheme places much increased emphasis on oral/aural skills (the only test of writing will be optional), and is concerned to make communicative effectiveness a significant criterion in assessing pupils' FL performance. In addition, substantial responsibility for the conduct of assessment (especially of the speaking test which is to account for 50 % of the total mark) will be devolved under the new scheme to the classroom teachers themselves.

THE RESEARCH PROJECT

The "School-Based Assessment of Communicative Foreign-Language Skills in S3 and S4 project was funded as part of a special research programme in parallel with the Munn/Dunning development programme, and runs from November 1983 to October 1986. The project has three main aims: a) to develop a set of criteria for the assessment of pupils' communicativie performance, which will be both theoretically well-founded, and usable by teachers for assessment of their pupils' FL proficiency, under "normal" classroom conditions; b) to develop quidelines for the consistent administration by teachers of oral interview tasks, having the main characteristics of "real life" communicative speech events; and c) to study the place of formative assessment in "communicative" instructional strategies at intermediate levels. These three areas are being investigated by means of an action research programme involving collaboration between the research team and teachers of French, German and Spanish working with S3/S4 pupils (i.e. 14 - 16 year-old pupils) in secondary schools in Central Scotland.

Teachers working in six schools have agreed to become involved in the research project, and as a preliminary to a series of focussed action research studies dealing with these assessment issues, a number of familiarisation visits to these schools have been carried out by members of the research team. The visits have involved interviews with all FL teachers in the schools concerned, plus unstructured observation of teaching and assessment from S2 to S4 (with a main focus on S3). For the

research team, these visits were a necessary preliminary to more focussed work on assessment issues; given the classroom based, collaborative nature of the research design, thorough familiarisation with the curriculum and methodology forming the current context for classroom assessment practices was essential. Moreover, they have also provided a unique record of the nature of FL teaching and testing in the middle years of secondary schooling, at a time when teachers are still preparing pupils for the traditional O Grade examinations, though with increasing consciousness of the impending changes in syllabus and certification procedures.

THIS PAPER

The main part of this paper will provide an account of the FL classroom assessment practices observed during these familiarisation visits, in the Spring 1984. These data are evaluated in terms of the extent to which the assessment practices currently in use in cooperating departments appear to elicit evidence about pupils' communicative FL proficiency. A description will then be given of teachers' views concerning the teaching and assessment of communicative skills, as collected from interview data. Finally a short report will be made of a series of group discussions held by members of the research team and cooperating teachers concerning a definition of communicative competence. The paper will conclude with a presentation of some ideas for the action research investigations.

COMMUNICATIVE COMPETENCE AND COMMUNICATIVE SPEECH EVENTS

The framework of FL communicative competence to which reference is being made in evaluating current practice at this stage of the project is that proposed by Canale (1983), a further development of that first put forward by Canale and Swain (1980). Canale indentifies four major components of communicative competence:

1. Grammatical competence (including phonology, orthography, vocabulary, word formation, sentence formation)

2. Sociolinguistic competence (expression and understanding of social meanings appropriate to different sociolinguistic contexts, and of grammatical forms appropriate for their expression)

3. Discourse competence (knowledge of different linguistic genres, together with their related devices for cohesion and coherence)

4. Strategic competence (ways of coping with grammatical, sociolinguistic, discourse and performance difficulties).

Acceptance of such a framework has implications for assessment as any assessment scheme or task must be such as to elicit performance data from the learner relevant to each different category of competence.

Littlewood (1977) and Mitchell and Johnston (1981) have tried to specify the characteristics of classroom speech events likely to provide "communicative" experience for the learner, in this expanded sense. Morrow (1981) has similarly attempted to indentify salient features of communicative assessment tasks. From these different proposals can be derived a number of more or less directly observable characteristics of communicative FL speech events (relevant to oral, face-to-face interaction):

1. Unpredictability of FL utterances (and the absence of structural constraints on what is said)

2. "Newness" of information being conveyed

3. "Natural" turntaking (participants initiate etc. as appropriate to their real or simulated conversation role)

4. Message-related rather than form-related feedback from the learner's interlocutor.

Reference will be made to these criteria in considering the effectiveness of current classroom-based assessment tasks in eliciting communicative performance data from S3 learners.

ORAL TEST OBSERVATION

As part of the initial familiarisation process, special visits were made to all six schools to observe examples of the S3 end-of-term tests currently in use. These tests had been developed within the departments using them and were mostly conducted by the class teacher. Eleven testing sessions were observed: ten of these involved the use of tests modelled on the current O Grade Oral Proficiency examinations which the pupils concerned will be sitting in the summer of 1985 - eight on the so-called "Traditional" O Grade examination, two on the so-called "Alternative" O Grade, which takes place in the language laboratory. The remaining test observed was with a class who were not preparing for the 16+ examination. As the observer was unable to take a note of the pupils' responses recorded in the language laboratory, and as the test with the "non-certificate" class provided only an isolated example, comments will be confined to the face-to-face interview tests which were modelled on the Traditional O Grade.

The eleven sessions included tests of French, German and Spanish. Seventeen teachers were observed in total: 15 testing 91 pupils in face-to-face interview and two administering tests to 55 pupils in the language laboratory.

The aim of these observations was for the research team to gain a picture of the testing procedures being used in the cooperating schools so as to provide a context in which we could then carry out the interventionist action research phase. The next part of this paper will give a description of the test observations

and look at how professionally and consistently teachers were operating and also at what in fact they were testing.

Test Format

The format of the test in four of the schools was modelled on the Traditional O Grade Oral Proficiency Tests. This test consists of four parts:

A. Reading Aloud
B. Answering Questions on a Picture
C. Answering General Questions
D. Role Playing

Each school had a selection of passages, pictures, general questions and role-play situations for use in the test but in each case it was the teacher/tester who chose which one the candidate should work from. There were, however, notable differences in the passages and the questions used by the various schools. The reading passages varied in length and type (i.e. whether prose or dialogue) and the range of tenses required to answer the questions went from the present tense only to present, past and future.

All but one of the departments used a role-play such as suggested in the SEB guidelines, providing testees with five printed instructions in English on a situation to be transposed by the pupils into the FL, the teacher being required to "respond appropriately". One department, with future developments in mind, provided pupils with three scenarios to prepare in advance. During the test, the teacher chose one and asked the pupil to engage in a free conversation, giving instructions such as "Do as much talking as you can", "This is a chance for you to show off". This was the only part of any test observed which was not discrete-point (excluding the Reading Aloud) and which lead to a more "natural" open-ended conversation.

Language of Test Management

Although all the teachers were following the same examination guidelines (which in fact make no explicit mention of the Language of Test Management), differences were noted in the approach of the various testers.

Six of the 17 teachers introduced and managed the test entirely in the FL. This was obviously seen by the testers concerned as comprising part of the test (even though this was not incorporated into the marking scheme.) Six testers used both L1 and the FL, tending to speak less English as the tests proceeded, thus providing varied experiences for the different testees. Five teachers introduced the test and gave all explanations in English. It was in fact this group which appeared the most concerned about putting the pupils at their ease and making quite sure they knew what they were supposed to be doing.

Scoring

The four departments preparing pupils for the Traditional "O" Grade on the whole used the SEB (Scottish Education Board) guidelines for marking, both in terms of the marks to be given (sections A, B, C, D to be marked each out of 15) and the criteria for awarding of marks (i.e. a series of "bands" to describe Reading Aloud, and three marks per utterance in the other three sections with stipulations in sections B and C as to the use of finite verbs). The present SEB guidelines concerning the role-playing stipulate that each of the five single pre-structured utterances are to be impression marked on the basis of communication:

> Full communication achieved, plus full accuracy and/or idiomatic fluency of expression. 3 Marks.
>
> Full Communication achieved. 2 Marks.
>
> If in doubt as to full communication. 1 Mark.
>
> No Communication. 0 Marks.

The only department which chose global rather than discrete-point marking for the role-play devised its own pegged marking system based on task completion, mentioning such characteristics of communicative speech acts as fluency and willingness to initiate.

All 17 teachers wrote the marks for the Reading Aloud either during the reading or as soon as the pupil had finished. For the Questions on a Picture and the General Questions, teachers tried on the whole to write a mark immediately after the utterance. However, on occasions, a teacher would fall behind and then have to fill in all the marks retrospectively. Moreover, in the case of the set role-play, four out of the five teachers observed found they were unhappy writing a mark and playing their part simultaneously, so they filled in the mark retrospectively, again despite the fact that they had five discrete utterances to score. In the department working with scenarios, a global mark was registered retrospectively (in any case, all teachers in this school audio-recorded this part of the test for retrospective consideration and departmental discussion of the marking).

Tester Behaviour

Wide variations were noted in the way different testers elicited responses and prompted the testees. The SEB guidelines state that for sections B and C a pupil may ask for the repetition of a question without specific penalty within the time limit. Very few pupils in fact asked for repetition; most gave no linguistic response when the question was not understood. However, faced with silence, the majority of teachers were willing to repeat or rephrase the question a number of times even using paralinguistic meas (for example, pointing to objects in the picture, or pointing to their watch to indicate time, out of the

window to indicate weather). It was in fact felt that some teachers were so keen to get an immediate reaction that they were giving the testee little opportunity to think. Only two of the less experienced teachers stuck rigidly to the set question, repeating it loudly in the case of pupil non-response. Feedback offered by the tester varied from silence (accompanied by an obtrusive writing down of the mark) to a friendly response which often involved a repetition of the pupil's utterance (without the errors), suggestion of a word the pupils had been unable to find and a lead-in to the next question. In such cases the writing of the marks tended to be less obvious.

The variations in teacher behaviour during the tests can perhaps be partly explained by the apparent lack, in some cases at least, of departmental discussion on the question of test administration prior to the test.

Although some teachers made very obvious efforts to treat pupils consistently, the general impression gained by the observers was that individual teachers did react differently to different pupils, not only in the choice of picture/questions/role-play (on no occassion, as has already been mentioned, was there any evidence of pupil choice) but also in the feedback they offered. Tiredness may well have been a factor here. However, the evidence to support these impressions is only fragmentary and more empirical work needs to be done.

Dimensions of communicative performance being tested

What dimensions, then, of foreign language performance, were being tested in the observed oral tests? On the whole, the only component of communicative performance being tested in any systematic way was grammatical performance. Guidelines for the marking of the questions in sections B and C (Questions on a Picture and General Questions) lay stress upon grammatical accuracy, urging testers to watch for "serious errors". The testee throughout the test, then, was being mainly scored on his/her ability to compose a correct sentence (or an appropriate subordinate clause) in the target language. There was little scope for the testee to display extended discourse competence, as interaction was limited almost entirely to single-sentence questions followed by single-sentence answers. Nor was there evidence of reward being given for proof of competence in strategic repair on the part of the pupil. Instances of the latter were in any case rare, with only the occasional pupil uttering phrases such as: "je ne comprends pas ce mot", or "Pouvez-vous répéter la question, s'il vous plait?", and interlocutor behaviour designed to elicit them was not built systematically into the test tasks. Furthermore, in the area of sociolinguistic competence little account was taken of appropriacy, with pupils, for example, in the role-playing addressing coach-drivers and shop keepers as "tu". This is consistent with the Traditional O Grade guidelines for the Oral Proficiency Test, which state that: "The mode of address (formal or informal) to which the candidates are accustomed will be used." In fact, in one of the test batteries, one set of general questions contained both "tu" and "vous"

forms, presumably to test whether the testee could comprehend both forms purely on a linguistic basis.

Only in the scenario-based role-play in the one department was there evidence of an attempt on the part of the testers to extend the boundaries of commucative performance being tested. As regards strategic repair, for example, the pupils, who had been told to try to keep the conversation going, showed willingness to use other words in place of the one they did not know or had forgotten (eg. one pupil said "No me gusta... la paella", when she wanted to say "la comida" and this was understood and picked up by the tester). A chance to cope with a longer, open-ended section of dialogue discourse was provided (the number of pupil utterances was in the region of 30 compared with 5 in the discrete-item test) as indeed was the opportunity to evince a basic knowledge of the rules of turn-taking. Appropriacy was also introduced into the test in the form of the different characters appearing in the role-plays (a young Spaniard whom pupils could address informally, and a waiter or a shop-keeper where the formal address was more appropriate). The open-ended nature of the role-play allowed for unpredictability, which was introduced on the part of the teacher when making an arrangement to meet someone (he announced he did not want to go to the beach because it was so boring) and also on the part of the pupil (who refused to buy the shoes the shop assistant was trying to press on her because she just did not like them). Finally "newness" of information was only really possible in an unscripted section of the test, therefore again in the scenario role-play. On the whole, then, one must conclude that only in this section of one of the tests was a wider definition of communicative competence in evidence.

Conclusion on Test Observations

The observations of the teaching and testing made by the research team seemed to reflect a broad consensus among teachers that what is teachable and testable in S3 is essentially grammatical competence in Canale's sense. In order to see how far this apparent consensus derived simply from the constraints of the present O Grade and how far teachers are in active agreement with this limited interpretation of communicative competence, it was decided to obtain a general survey of current teachers' views on curriculum and assessment by conducting interviews with the teachers in the cooperating departments.

INTERVIEWS

All teachers were approached and almost all agreed to be interviewed, making the total number 37. The transcripts have been analysed systematically, using a simple procedure of content analysis. The next section of this paper will give a general description of teachers' views on current patterns and future developments of FL teaching and assessment, with special reference to communication skills.

The first striking point to emerge from the interviews was the predominance of the present O Grade examination in teacher thinking and teachers freely admitted to this. Their main aim at the present - and this forms the basis of their teaching and testing - is to enable as many pupils as possible to get through the 16+ examination. When asked, however, it they were satisfied with the examination, strong discontent was expressed, not only with the various sections of the test but also with its stress on formal accuracy rather than on the expression of meaning, and especially with its lack of concern with oral expression.

Indeed, the second area of consensus emerging from the interviews was the commitment of the teachers to oral work. A large number of them stated that the aim of teaching a foreign language should be "communication", which they defined in terms, for example, of "beign able to get over what you want to say and understanding what someone says to you", or "the ability to convey and receive information in the language". Far from stressing formal accuracy teachers spoke time and again of getting across the basic message, of the passing over of "real information". They did, however, anchor their definition in a situation, stating that at this level they would like their pupils to be able to "cope with" the sort of "survival" situation a tourist might meet with in the country where the language is spoken, although some teachers did add that they would also like their pupils to be able to talk about themselves and their own lives. In addition, teachers alluded to communicative speech events such as natural turntaking (and pupil initiation) and unpredictability.

When asked their views on assessment, teachers came down heavily on the side of some open-ended role-play or conversation, which they felt could test the various facets of communication they had spoken of. They were critical of the very discrete-point test they had been observed conducting. Another point which many of them made was that part of the oral assessment should be based on observation of pupils taking part in classroom communicative activities (some used the term "eaves-dropping"), and that a formal 15 minute test could not stand on its own as a valid test of a pupil's communicative ability.

Teachers, then, showed themselves to be well aware of the need to change assessment tasks and procedures if a reliable and valid test of pupils' communication skills was to be attempted. Indeed a lot of what they were moving towards in their interview responses appeared later in the Joint Working Party Report (June 1984). This has recommended that the oral mark be made up of a "series of speaking activities or tests which arise out of normal class work" (in S4) and a "formal test of speaking" based on the completion of various tasks, in which pupils are to be assessed on their "ability to tackle the development of the tasks and any associated conversational acitivities". So, even if the teachers interviewed are not yet clear as to how to integrate their own suggestion into tests which set out to assess communicative skills, it would appear that the majority of them are in tune with the direction the reforms at 16+ are taking.

DISCUSSION OF COMMUNICATIVE COMPETENCE FRAMEWORK WITH THE TEACHERS

The next phase of the project involved discussing Canale's framework of communicative competence with the departments in the various schools, to see whether teachers accepted any or all of the components and whether they had additional suggestions to make. Up to the time of the writing of this paper, four schools have taken part in such discussions and the following section gives a brief report of these.

The general consensus on the subject of grammatical competence was that it was an important component but one that had beed over-stressed in language teaching in the past. Explicit grammar teaching had its place but should come after the pupils had learnt to use the language. A number of teachers felt that those aspects of grammar which play a vital part in the communication process should take precedence over grammatical accuracy for its own sake.

Whereas teachers generally agreed that sociolinguistic competence was a vital component of communicative competence especially when a learner was in the foreign country, they were very hesitant about its playing too large a role in the teaching and especially the assessment process. The consensus was that at this stage, most pupils had too many other aspects of communication to concentrate on. Better to leave it for more advanced learners.

Discourse competence was a component that teachers again agreed was an important one, and one that they are becoming more aware of. The majority view, however, was that it was not so important at basic level and could be used as a means of differentiation between various levels of linguistic competence.

Strategic competence received general acclaim as to its importance in terms of keeping the conversation/communication going. It was also an area teachers feel could and should be integrated into the teaching and testing arrangements.

One other possible dimension of communicative competence (not included by Canale in his framework) which was discussed with the teachers was that of confidence and willingness to speak. The teachers on the whole spoke out strongly against integrating such aspect into the assesment scheme, while agreeing that they try to further them in the teaching process. However the Report of the Joint Working Party puts "Willingness to communicate" as a feature to be taken into account alongside "Language Used" in a consideration of "Differentiation by Candidate Performance". It states:

"A high level of performance will not be achieved without an extensive knowledge of the language, but factors other than such knowledge will be taken into account when evaluating candidates' performance: willingness to attempt to communicate, to interact in conversation, to take the initiative, to go beyond

uniformly minimal responses, to accept some responsibility for sustaining the dialogue".

Although we have seen that teachers are keen for their pupils to learn to interact and take the initiative, they do share fears as to how the more shy candidate is going to fare under the new system.

On the whole, then, teachers have shown themselves to be well aware of the various components of communicative competence as suggested by Canale, and that they accept the role of each of them, the only reservations being with regard to what can be expected from the level of learner being considered by the project (S3 and S4).

CONCLUSION - THE NEXT PHASE OF THE PROJECT

The data collected during the test observation and the interviews and discussions with teachers provides much material on which to base the next phase of the research project - the series of small-scale action research studies for which researchers and classroom teachers will be jointly responsible. Possible areas of investigation and ones in which teachers have already expressed an interest would be, for example, the validity of the global conversation task, tester/interlocutor behaviour, inter-tester and intra-tester reliability and the scoring procedure. The project team look forward to teachers continuing to contribute ideas and suggestions concerning the identification of assessment needs of their classess and also concerning monitoring and cooperative evaluation of the tests they implement.

REFERENCES

Canale, M. (1983) "From communicative competence to communicative language pedagogy". In Richards, J.C. and Schmidt, R.W. (Eds) LANGUAGE AND COMMUNICATION. Harlow, Essex: Longman, pp. 2 - 28.

Canale, M. and Swain, M. (1980) "Theoretical bases of communicative approaches to second language teaching and testing." JOURNAL OF APPLIED LINGUISTICS, 1.1, 1 - 47.

The Dunning Report (1977) "ASSESSMENT FOR ALL. Report of the Committee to review Assessment in the Third and Fourth Years of Secondary Education in Scotland". Scottish Education Department.

Joint Working Party on Modern Languages. (1984) "REPORT ON FRENCH AT FOUNDATION, GENERAL AND CREDIT LEVELS". Scottish Examinations Board.

Littlewood, W.T. (1977) "Defining 'communication' in foreign language teaching". LINGUISTISCHE BERICHTE 52, pp. 83 - 92. Stirling, Mimeo.

The Munn Report. (1977) "THE STRUCTURE OF THE CURRICULUM IN THE THIRD AND FOURTH YEARS OF THE SCOTTISH SECONDARY SCHOOL", Scottish Education Department.

Morrow, K. (1981) "Communicative language testing: Revolution or evolution." In Alderson, J.C. and Hughes, A.H. (Eds), ISSUES IN LANGUAGE TESTING: ELT DOCUMENT 111. London: The British Council, pp. 9 - 25.

REFLECTIONS ON THE CONTENT VALIDITY OF HOLISTIC TESTS

Christine Klein-Braley (University of Duisburg)

Probably the test forms which cause the language tester the most headaches are those which need no justification whatsoever to the naive user: the essay, the translation, the précis, the written reproduction of an aurally presented text etc. In what Spolsky calls the traditional trend in language testing, such test procedures are the norm. Just as the budding pianist has to learn to read the music and practise scales in order to produce the final integrated performance of playing a sonata, so too such test forms offer the examinee the opportunity to demonstrate a synthesis of all that has been learned in a "facile integrated performance" (Carroll, 1961).

These traditional tests have high face validity. But I think that ultimately the validity concept underlying essays, translations etc. is that of content validity: I imagine that the majority of supporters of traditional tests would find it easy to agree with a statement like "holistic tests sample the learner's use of all elements of language simultaneously in an integrated fashion".

Content validity

The APA test standards state:

> Evidence of content validity is required when the test user wishes to estimate how an individual performs in the universe of situations the test is intended to represent. Content validity is most commonly evaluated for tests of skill or knowledge. (APA, 1974:26)

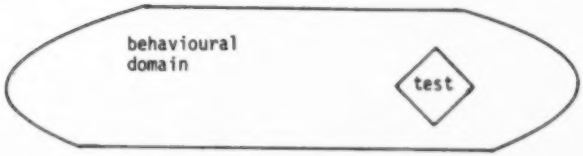
and continue:

> To demonstrate the content validity of a set of test scores, one must show that the behaviours demonstrated in testing constitute a representative sample of behaviors to be exhibited in a desired performance domain. Definitions of the performance domain, the user's objectives, and the method of sampling are crititical to claims of content validity. An investigations of content validity recuires that the test developer or test user specify his objectives and carefully define the performance domain in light of those objectives. The definition should ordinarily specify the results of learning rather than the processes by which learning is either acquired or demonstrated. It should be sufficiently detailed and organized to show the degree to which component tasks make up the total domain. (APA, 1974:28)

Graphically, we can represent this type of validity by Figure 1. The test is a representative sample of the behavioural domain, and therefore conclusions can be drawn from test performance to real-life performance.

Figure 1

Concept of content validity



The critical terms in the APA definition are:

- definitions of the performance domain
- the user's objectives
- the method of sampling.

Here I should like to discuss the implications of modern approaches towards content validity for traditional language tests.

Approaches to content validity

It is reasonable easy to ensure content validity when the domain to be sampled is finite and when all possible test items in the domain can be listed, even if only theoretically. When I was teaching infants our teaching objective in number skills at the end of the first year was the "number bonds to 2o". This meant that a child had to be able to add, subtract, multiply and divide any two numbers to obtain a third, but that no term in the operations was to be bigger than 2o. You can see that it is very easy to determine whether any given item belongs to the domain or not, and also that it is possible to list every item which belongs to the domain. However this would be rather a long list, so it is easier to write a rule which will generate suitable items. This might run:

$$a \begin{matrix} + \\ - \\ \times \\ \div \end{matrix} b = c \text{ where } a, b, c \leq 2o$$

The test would then consist of a random sample of such items, where the terms a & b are generated randomly. From test performance we could draw generalisations about whether a child had mastered the number bonds to 2o or not.

However there is an additional problem. The items have to be made

into test items, i.e. one of the terms a, b, or c has to be removed so that the child can demonstrate his or her mastery of the operations by inserting the correct solution. My experience of teaching infants tells me, that in fact the items are easier if term c is omitted

i.e. $4 \times 2 = ?$

than if a or b is omitted:

i.e. $? \times 2 = 8$
$4 \times ? = 8$

It is easy to see why this is the case. In the last two items the child has to work out that he or she must perform not a multiplication operation as the arithmetic operator implies, but a division if the correct solution is to be found. So omission of a or b demands a higher level of understanding of the operations concerned than omission of a. In other words, it is not only the representative sampling of the domain which is important, but also the types of item involved. This means that test method too is an integral part of the total test procedure.

The simplest technique for ensuring content validity is:

1. to define the domain either by listing the items or by devising rules which generate only items which belong to the domain;
2. to take a random sample from the domain;
3. to select a suitable item form for the testing.

Obviously, as in all tests, the items should be objective and equally obviously more than one item should be used in order to minimise measuring error. The final test is valid as a result of the construction procedures: content validity is built in right from the start.

In most cases the problem is not as easy to solve as in our example. For instance the domain can consist of an infinite number of items, or it can be rather difficult to determine where the domain boundaries are since it, itself, is infinite. For this reason it has become usual nowadays to construct tests intended to have content validity by means of what is generally called a Tyler matrix.

The Tyler matrix

The Tyler matrix is a two dimensional matrix with behavioural classes on one axis and content classes on the other:

For our example, the Tyler matrix would look like this:

Figure 2

Tyler matrix

CONTENT

Behaviour	Solve a op b = ?	Solve ? op b = c Solve a op ? = c
+		
-		
x		
÷		

The behavioural classes are usually hierarchically ordered, so that satisfactory performance in a higher-order category automatically includes the lower-order category.

Content validity and psychometric-structuralist tests

If we now turn our attention to one of the language testing classics, Harris 1969, we find that language testing too has long been using such matrices:

Figure 3

Language test construction matrix

	Language Skills			
Components	Listening	Speaking	Reading	Writing
Phonology/ Orthography				
Structure				
Vocabulary				
Rate & gen. fluency				

(Harris, 1969 : 1oo)

You see, therefore, that psychometric-structuralist language tests too are based on the concept of content validity.

If we take the Harris matrix as a basis for test construction , it is easy to see that in fact it is a supermatrix. Each of its cells, or each row can be enlarged into a separate matrix. For instance, we might choose to select the row STRUCTURE, and to subdivide this into a further set of content classes:

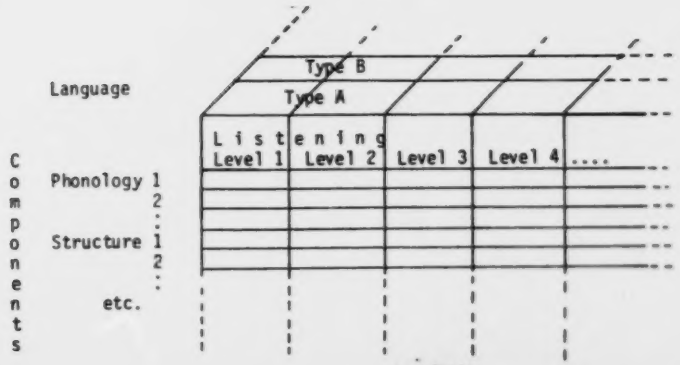
S T R U C T U R E

Determiners
Prepositions
Tenses
Active / Passive
.
.
.

or to subdivide LISTENING into categories which begin with the relatively easy listening task of a text, presented straight, read with "teaching pronounciation" so that every word can be clearly heard, with no hesitations, back tracks or "erms", and no background noise and which end with the most difficult task: listening to a genuine, unscripted conversation between native speakers, with all the "erms", interruptions, structural changes in mid-sentence and background noise that is typical for real-life communication. These categories or levels of difficulty are clearly hierarchical in a way that the original Harris categories are not.

The content classes in this case could either be those of the original Harris matrix, maybe broken down into subcategories, or we could choose to use a sampling procedure for all the various types of text content which are possible: a lecture on engineering, a scripted discussion on women's lib., an unscripted family discussion on where to spend the holidays this year. These must be regarded only as samples, as examples of possible subject matter, because one could just easily choose a lecture on extra-sensory perception, a scripted discussion on whether the 16-year-old daughter should be allowed to go to a pyjama party or not. Or we could try to include both Harris's components and the various topic areas in the sampling scheme. This would give us a three-dimensional matrix:

Figure 4



This too, is a familiar matrix to language testers: it is the one proposed by Cooper in 1968.

We still have the question of test method, i.e. item construction. We can test listening comprehension by using individual open-ended or multiple-choice items, we can ask the examinee to make deductions from what he or she has heard, to give us a brief summary. This is rather difficult to represent graphically, and so we use the mapping sentence (cf. Klauer, 1984). For our listening comprehension matrix the sentence would state something like:

The examinee demonstrates listening comprehension ... by correctly solving	ITEMS Type 1 . . . n	..which test..	COMPONENTS Phonology 1 . . . n 1 . . . n	of..	DISCOURSE Type 1 2 . . n	....
					LISTENING Level 1 2 . . n	at ...

If we want to make deductions about the entire matrix we can either sample each cell once, which would give us

$$n_{(\text{items})} \times n_{(\text{components})} \times n_{(\text{discourse})} \times n_{(\text{listening levels})}$$

or we can take a random sample of the cells. Of course, if we want more information about any specific cell, we have to take a number of items from that cell. As you can see, test length is rapidly becoming infinite at this point.

There is another problem too. So far I have assumed that it is fairly easy to specify the substructures of the various dimensions, but, as we all know, this has not yet proved to be possible. Both the structuralists and the Chomskyans made brave attempts, but both ignored important features of the language: the structuralists its creativity, the Chomskyans its functionality. So the components are going to cause us headaches. So far as discourse topics are concerned, the number is probably infinite, and the same goes for levels of listening comprehension. We are going to have to sample here, but we have no way of telling whether our samples are representative or not. In addition we must consider the question of weighting. Is each of the cells equally important in the total activity of listening comprehension, or are some more important than others, and if so, which?

These were, of course, among the reasons why in the early seventies linguists became dissatisfied with psychometric-structuralist testing procedures. Other reasons included the problem of relating test beha-

viour to real-life language performance, i.e. of predicting the ability to function linguistically under real-time constraints from tests whose artificiality was only too apparent. Yet another reason was provided by the statistics emerging from psychometric-structuralist tests: the subtests developed for such allegedly distinct skills as listening and reading comprehension, grammar and vocabulary, all correlated so highly with each other that there seemed good reason to assume that, in fact, language was a unidimensional entity, and that language proficiency had to be viewed as holistic.

As a result of the apparent impossibility of ensuring content validity in psychometric-structuralist tests new test forms were proposed, among others dictations, cloze tests and later C-Tests. What about the content validity of these tests?

Content validity and post-modern tests

Post-modern tests such as cloze tests, dictations and C-Tests are all based on the theory that language is a holistic entity, and that the same type of language processing occurs whether the activity is "active" or "passive". For this reason the behavioural dimension of the matrix can be collapsed into one single column, headed "all". The former content specification also collapses into one dimension, because now the behaviour involves the use of all elements specified here as part of the language processing activity. This means that the content dimension can now be specified as "discourse topic: any", because post-modern tests make the assumption of text equivalence.

Figure 5

Content	Behaviour
	Language processing (all elements)
discourse topic: any	POST-MODERN TESTS

Obviously the same arguments can be advanced for all the traditional tests: in the essay we also have the assumption that all aspects of language processing are involved, and that the discourse topic is irrelevant. The same applies to such tasks as translations, summaries, interviews, and so on. Thus, if cloze tests, C-Tests, and dictations are content valid according to this approach, so too are the traditional test forms.

However, we also have the third dimension of the matrix, namely test method, and here there are major differences.

Text selection

So far as test construction is concerned, virtually all methods begin with text selection: one text. The text is viewed as a random sample of the universe of texts. However, as Nunnally points out (1967), every test item has a component of specifity: characteristics which are unique to any particular item. Item specifity is undesirable,

because in any test it is not performance on the individual item which interests us; the item is seen as one of many possible items, i.e. again it is a sample. The more item specifity an item has, the less suitable it is as a test item. In holistic tests the use of only one text means that item specifity can play a much larger role than is desirable. Item specifity is at its most problematical in the essay, of course, because the instruction "Write an essay of 300 words on the following topic..." is a direct invitation to the examinee to produce his or her own text. The 5 or 6 texts of the C-Test are a much better approximation to a sampling process.

Test production

This stage is only relevant for post-modern tests. While the traditional tests simply take the selected text and ask the examinee to perform according to the instructions given, post-modern tests use the concept of random sampling again to delete segments of the text for examinee processing. In cloze tests this is the "nth word" deletion principle, in C-Tests the "rule of 2". In both cases the systematic deletions are intended as an approximation of random sampling. As I have shown, "nth word" deletion is not entirely satisfactory (Klein-Braley 1981), while the "rule of 2" does delete representative elements of the text. But the intention of sampling is present in both cases.

In the dictation random sampling is not necessary because here the examinee is confronted with the entire text and is asked to convert the aural input into a graphic representation.

The importance of this step lies in its effect on the scoring procedures. In all the post-modern tests there is a correct solution to the test item involved. Admittedly one can choose whether to count spelling errors or not, but in dictations it is usual to deduct one point for every wrong word, in cloze and C-Tests to add one point for every correct restoration. In cloze tests we have the added problem of whether to use exact scoring or acceptable scoring. Exact scoring is entirely objective; acceptable scoring can be made objective if the examiners are prepared to spend sufficient time over it.

Traditional tests, on the other hand, are not objective. In translations we have the notorious problem of deciding on which type of fidelity we are aiming for. Often this is phrased as "as free as necessary but as accurate as possible. So it is quite conceivable that two scorers would disagree entirely on the evaluation of an examinee's rendering of the text, even though it was linguistically correct.

Once again the essay is the most problematical test form, since not only do we have the problem of scoring examinee production, we even have to devise some technique for bringing entirely different texts (those produced by the individual examinees) into the same scale of measurement. Consider the two extremes: one examinee writes very short, very simple sentences, all of which are linguistically correct. A second examinee is more ambitious, writes more complex sentences, but misses the target some of the time. Who gets the higher mark?

The important of the test production techniques in post-modern tests is that they ensure objectivity at the same time as performing sampling.

Text (test) processing

In working through the text the examinee presents the examiner with a sample of his language. In post-modern tests this sample is determined by the text. Avoidance behaviour is not possible; the examinee must cope with what is set in front of him. This means that we are more likely to get a representative sample of the examinee's performance than in the traditional test forms where within broad limits the examinee decides what he or she will offer to the examiner. This is most obvious in the essay: both what the examinee says and the way in which he or she says it are a matter of free choice, and it would be a particularly stupid examinee who deliberately chose to use structures known to cause problems or to venture into content areas where the vocabulary is not in the repertoire. This also applies to the rather more structured tasks such as translation and prècis where the source text controls the content but the linguistic formulations are left to the examinee. In fact, a large proportion of my teaching time is devoted to showing examination candidates how to "get round" problem areas in their examinations.

Test interpretation

Because they are objective, post-modern tests provide us with a numerical score at the end of the test procedure. This performance score is considered to be a numerical representation of the examinee's competence. This assumption is legitimate, because we have used sampling processes to produce the test. Naturally we need to take factors such as test reliability and the error of measurement into consideration, but in general we can assume that the examinee which the higher score is actually "better" at the language, i.e. his or her competence is more highly developed.

It is very difficult to make such assumptions on the basis of traditional tests, since the process of converting the scorer's evaluations of examinee performance into a numerical scale is a very tricky one for obvious reasons. This means that test interpretation is very difficult.

Conclusions

Traditional tests (or what Spolsky formerly called prescientific language tests) and post-modern tests both take the same approach towards content validity. The examinee is supposed to show in an "integrated facile performance" that he can use the language fluently.

Where traditional tests fail is in the third dimension of the matrix: test method. The post-modern tests present the examinees with a sample or samples of the language and expect them to demonstrate their control of the samples. By doing this they have a reasonable chance of obtaining a representative sample of examinee performance. Traditional tests, however, allow examinees to select their own samples, and so there is no guarantee that the test scripts are representative.

If it is considered necessary to use traditional tests, then there are a number of recommendations which can be made. They all involve more work, but this is necessary in order to tighten up the test process:

1. More than one text should be used (text sampling)
2. Examinees should be provided with much stricter guidelines about the type of performance expected (test construction and text processing)
3. Scoring procedures should be specified in advance and techniques for converting assessment into numbers should be developed, together with guidelines for interpretation (test interpretation).

In view of the additional effort these suggestions would involve, test users should consider whether a post-modern test would not serve their purposes just as well as a traditional test since so far as their approaches to content validity are concerned there are no differences.

Note

I have been greatly helped in the writing of this paper by my discussions with Ulrich Raatz. Any errors are, however, all my own work.

Bibliography

APA: American Psychological Association, 1974. Standards for educational and psychological tests. Washington: APA.

Carroll, J.B. 1961. "Fundamental considerations in testing for English language proficiency of foreign students." In Testing the English proficiency of foreign students. Washington: Center for Applied Linguistics. 30-40.

Cooper, R.L. 1968. "An elaborated language testing model." In Brown, H.D. (ed) 1968. Language Learning Special Issue No. 2: 57-72.

Harris, D.P. 1969. Testing English as a second language. New York: McGraw-Hill.

Klauer, K.-J. 1984. "Kontentvalidität." Diagnostica 30, 1: 1-23.

Nunnally, J.C. 1967. Psychometric theory. New York: McGraw-Hill.

A PERCENTILE REGRESSION MODEL FOR THE NUMBER OF ERRORS IN GROUP CONVERSATION TESTS

Erkki P. Liski (University of Tampere and the Academy of Finland)
Simo Puntanen (University of Tampere)

Introduction

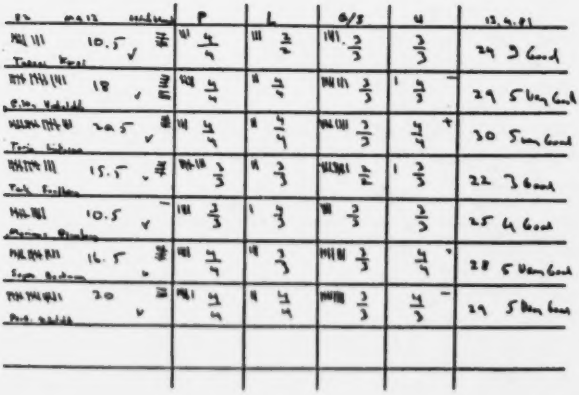
In this paper we put forward a statistical model for analysing the results of group conversation tests in spoken English. The variation of the number of errors was explained by the number of utterances spoken by a testee. Our data were obtained from the results of group conversation tests carried out at the University of Tampere from 1977 up to 1981. The number of students under consideration was 639. On the basis of these same data we analysed in another study (Liski and Puntanen 1978) the language performance of Finnish testees in relation to certain background variables like sex, school results etc. This earlier paper contains a thorough statistical description of our data.

In this study estimates of percentile curves for the number of errors are of greater interest than the mean regression line. In our situation the dependent variable turns out to be a compound Poisson variable. Explicit parametric assumptions are made using a linear model for scale and heterogeneous variance. We apply here the distribution theory described in more detail in two earlier papers (Liski 1984), Liski and Puntanen (1985). Also more far-reaching statistical discussion of the model can be found in the above papers.

1 Data

One day we happened to come upon Mr. John Clarkson, an English teacher in the University of Tampere, carrying some sheets full of figures and because statisticians are incorrigibly dedicated to finding facts from figures - whenever they meet figures - we asked what these sheets contained. He kindly told us; some consequences can be seen from the following pages.

John Clarkson's sheets looked like for example the followong:



The sheet above is a result of group conversation tests in spoken English. At first it struck us as somewhat surprising that a student's conversation ability could be measured by a quantitative method. Now that the problem field has become familiar, we see that there are indeed good possibilities for measuring such an ability, which is no doubt very important in life today.

As one may see, the examination sheet is full of figures implying that a kind quantitative assesment of the testee's ability to Speak English is applied. This testing system has been developed by Folland and Robertson at the University of Tampere, and it is described in Folland and Robertson (1976). In this section the reader is introduced to the data based on the examination sheets. To make the nature of data clear a brief description of the testing method will be next given.

Groups of students are tested in free discussion by a native English teacher for a minimum of five minutes per student. The typical group size is six, giving a 30-minute test for a group. At the beginning of the test a prerecorded tape is introduced for

example as being a report to an international study group, after which the main points should be freely discussed in English. The examiners take no part in the discussion, the contents and development of which lie entirely with the students. During the previous terms the students have had considerable practice in this.

A testee's perfomance is assessed on a marking sheet under four catagories: *pronunciation, lexis, grammatical structures,* and *use*. The number of *major* and *minor* errors under the four categories are noted, together with *plus points* under use and lexis. Also *the number of utterances* each testee produces during the test is noted on the marking sheet, where one utterance (roughly estimated) is at least ten words and a part utterance less than ten words. At the end of the test the number of errors and pluses in each category per ten utterances for each testee is calculated, and these figures are then applied to an errors-points scale to arrive at the final mark.

Folland and Robertson transcribed a number of recorded tests and examined the errors which had occurred, devising definitions for them. An error occurs where *the speaker fails to follow the pattern or manner of speech of educated people in use in English-speaking countries today*. Taken statistically major errors were very seldom made.

On the basis of discussions with the examiners, we have combined the number of minor and major errors so that one major error corresponds to two minors (except in the pronunciation category, where one major corresponds to three minors) for the purpose of this study. The figure thus obtained is called simply the number of errors, thus excluding the major-minor division.

The main variables to be considered in this paper are

TFREQ = the total *frequency* of utterances given by a testee,

EFR = the total *error frequency* of a testee.

The examiners, however, consider errors separately in different categories which gives a more versatile picture of the

testee's ability. In our data the variable EFR is simply the number of all errors made by the testee during the test.

It is obvious that the more the testee speaks the larger is the number of his errors. Hence the number or utterances should somehow be taken into account when assessing the testee's ability. As was earlier mentioned, the examiners simply consider the number of errors per ten utterances and then use their errors-points scale to arrive at the final mark. The errors-point scale is based on the examiners' *experience* of the "tolerable" number of errors. One simple *statistical* approach to the errors-points scale would be to consider the *conditional percentiles* of the number of errors (EFR) under different values of utterances (TFREQ).

The main interest of this paper lies between the dependence of the variables TFREQ and EFR, whose scatter diagram is shown in Figure 1 and frequency distributions in Figures 2 and 3. The objective of the study is to arrive at a statistical model of the data and particularly to estimate percentile curves for the variable EFR.

Our data were obtained from the results of group conversation tests carried out in the Universtiy of Tampere by John Clarkson from 1977 up to 1981. Also certain background variables such as sex and matriculation results were observed but they will not be considered in this paper. The number of students under consideration was 639. Originally the data consisted of 826 students but to reduce the heterogeneity of the students' background only students with the matriculation examination were included in the research. The examiners have found that there must be at least six utterances in order to assess the testee's ability in spoken English. On the other hand, ten utterances have always proved to be a sufficient sample size. Hence testees speaking less than six utterances are not included in the data here, thus the variable TFREQ varies from 6 to 40 in these data.

Figure 1

Scatter diagram between TFREQ and the number of errors

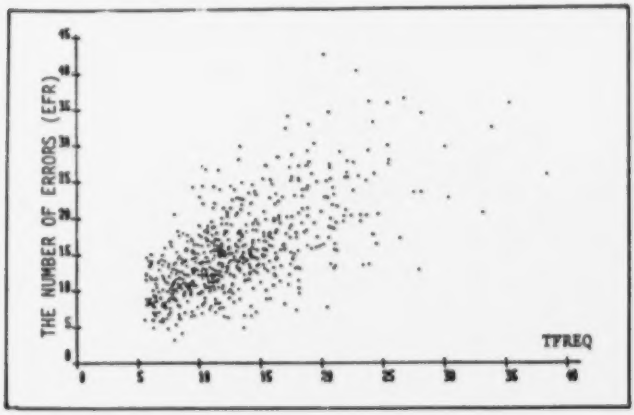


Figure 2

Distribution of the number of utterances (TFREQ)

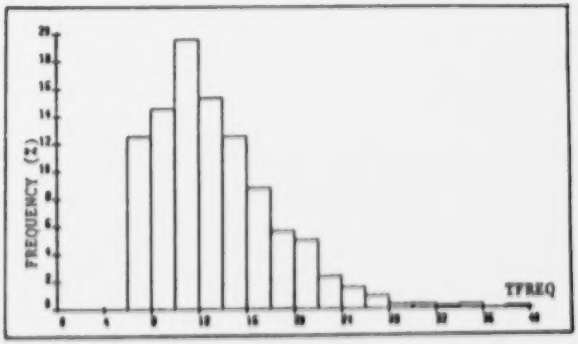
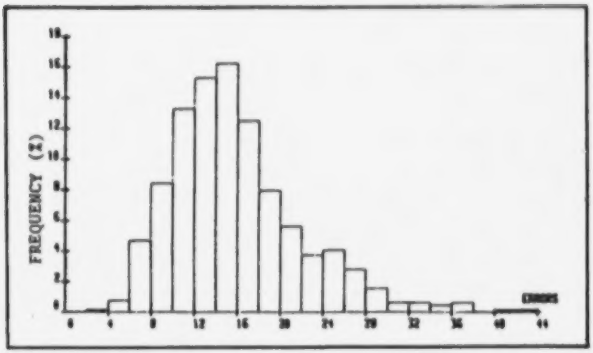


Figure 3

Distribution of the number of errors (EFR)



2 Modelling

As we explained in the previous section, testees spoke from 6 to 40 utterances during a test. The total number or errors made by a student varies from 3 to 43. In the whole sample the average number of errors per one utterance is 1.23. Let us consider first a particular student, who has spoken x utterances (in our sample $6 \leq x \leq 40$). Denote by y_j the number of errors made by a testee in the jth utterance he has spoken. Therefore, the set of utterances produced by a student can be characterized as follows:

utterance	1	2	...	x	the total number of errors
the number of errors in one utterance	y_1	y_2	...	y_x	$y = \sum_{j=1}^{x} y_j$

We assume that the number of errors in one utterance follows a Poisson distribution with the mean λ_x, where x is the number of utterances spoken by a student. Therefore we write

$$y_j \sim \text{Poisson}(\lambda_x), \text{ for all } j = 1,2,\ldots, x. \qquad (2.1)$$

Further, we assume that $y_1, y_2, \ldots, y_x$ are stochastically independent. As a sum of Poisson variables the total number of errors $y = y_1 + y_2 + \ldots + y_x$ made by a particular testee is also a Poisson variable:

$$y \sim \text{Poisson}(\lambda_x x). \qquad (2.2)$$

The value of the intensity parameter λ_x describes a student's proneness to errors. A high value of λ_x tells that a student's language ability is bad.

However, it is quite evident that proneness to errors, and hence the value of the intensity parameter λ_x varies from one student to another. Every particular student has a value of λ_x characteristic to him. Therefore, intensity parameter λ_x is a random variable. We have found (Liski and Puntanen 1983), on the basis of an empirical investigation that students who spoke seldom were more prone to errors than more talkative testees. Therefore we assume that the mean of λ_x depends on TFREQ(=x), and we denote $E\lambda_x = \mu_x$. Further, suppose that the variance of variable λ_x is σ^2, and $\text{var}(\lambda_x) = \sigma^2$ does not depend on x.

As was stated before, the mean μ_x decreases as x (the number of utterances) increases. In other, words talkative students tend to be better than students who speak seldom.

Taking the intensity parameter as a random variable yields the compound Poisson distribution (see e.g. Johnson & Kotz 1969). In this case the expectation and variance of (2.2) are

$$Ey = \underset{\lambda_x}{E}[E(y|\lambda_x)] = \mu_x x \qquad (2.3)$$

and

$$\operatorname{var} y = \underset{\lambda_x}{E}[\operatorname{var}(y|\lambda_x)] + \underset{\lambda_x}{\operatorname{var}}[E(y|\lambda_x)]$$

$$= \mu_x x + \sigma^2 x^2 , \qquad (2.4)$$

where x is the number of utterances (see e.g. Rao 1973, p. 97).

However, the number of errors per utterance (EFR/TFREQ) is used instead of EFR when assessing the student's ability in pronunciation. These figures are then applied to an errors-points scale to arrive at the final mark in pronunciation. As a direct consequence of (2.3) and (2.4) we obtain

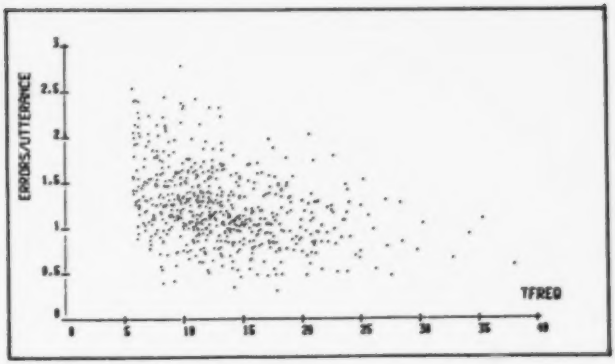
$$E(y/x) = \mu_x , \qquad (2.5)$$

$$\operatorname{var}(y/x) = \mu_x/x + \sigma^2 . \qquad (2.6)$$

Thus we find that also the variance of y/x depends on x.

Figure 4

Scatter diagram between TFREQ and the number of errors per one utterance



3 Regression

We assume a linear relationship between y and x. That is

$$y = \alpha + \beta x + \varepsilon \,, \tag{3.1}$$

where

$$E\varepsilon = 0 \tag{3.2a}$$

and

$$\mathrm{var}\ \varepsilon = \mu_x x + \sigma^2 x^2 \,.$$

However, we are more interested in estimating different percentile curves as a function of x than in estimates of regression parameters α and β. Since final marks can be determined on the

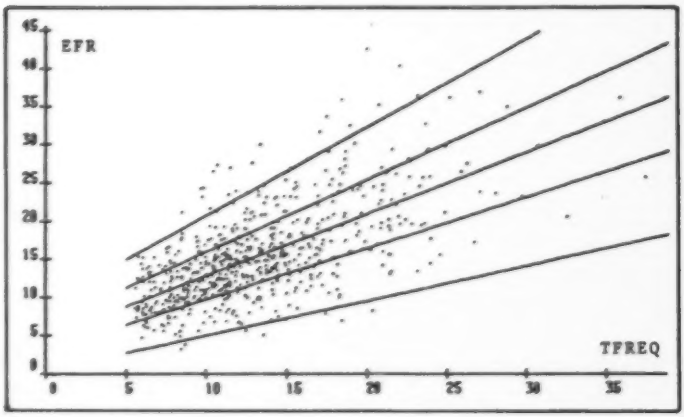
basis of the figures y, the examiners simply need various percentiles as a function of x. The percentile curves are useful both in constructing errors-point scales and generally in comparison of error frequencies with different values of x.

We estimate parameters α, β and σ^2 by the method of maximum likelihood, when the distribution of ε was assumed to be normal. The estimates were found to be

$$\hat{\alpha} = 4.788 \text{ (sd. } 0.481),$$
$$\hat{\beta} = 0.806 \text{ (sd. } 0.039),$$
$$\hat{\sigma}^2 = 0.032 \text{ (sd. } 0.007).$$

Figure 5

Estimates of the 5th, 25th, 50th, 75th and 90th percentile curves for the regression of EFR on TFREQ.



The estimated median line is

$$EFR_{0.50} = 4.778 + 0.806 \text{ TFREQ}$$

and the variance of EFR can be expressed as a function of TFREQ

$$var(EFR) = 4.778 + 0.806\ TFREQ + 0.032\ TFREQ^2.$$

The estimate of the 100 pth percentile curves for the regression of EFR on TFREQ were found to be of the form

$$EFR_p = 4.778 + 0.806\ TFREQ + (4.778 + 0.806\ TFREQ + 0.032\ TFREQ^2)^{1/2} Z_p,$$

where Z_p is the 100 pth percentile of the standardised normal distribution. As can be seen from Figure 5 percentile curves are rather linear.

When assessing a student's language ability, we usually consider the number of errors per utterance (u = EFR/TFREQ). Percentile curves for this variable are of the form

$$u_p = 0.806 + (4.778/TFREQ) + (0.032 + 0.806/TFREQ + 4.778/TFREQ^2)^{1/2} Z_p.$$

We see from Figure 6 that proneness to errors decreases when TFREQ increases. Thus more talkative testees are better than the testees who rarely speak. On the other hand, when the value of EFR/TFREQ is low (the best students) proneness to errors does not seem to depend on TFREQ (the 5th percentile curve). When TFREQ increases, the value of $u_{0.50}$ approaches 0.806 and the standard deviation of u approaches $0.032^{1/2} = 0.179$.

4 Discussion

We clearly perceive that the distributions of EFR and of the relative number of errors (EFR/TFREQ) depend on TFREQ. Although talkative testees make more errors than testees who rarely speak, proneness to errors decreases when TFREQ increases. Thus talkative students do better in this conversation test than testees who seldom speak. It also proves out that male students make more errors than female students, but there is almost no difference in the relative number of errors between boys and girls (Figure 6). In fact, talkative boys were less prone to errors than talkative girls.

Figure 6

Estimates of the 5th, 25th, 50th, 75th and 95th percentile curves for the regression of EFR/TFREQ on TFREQ.

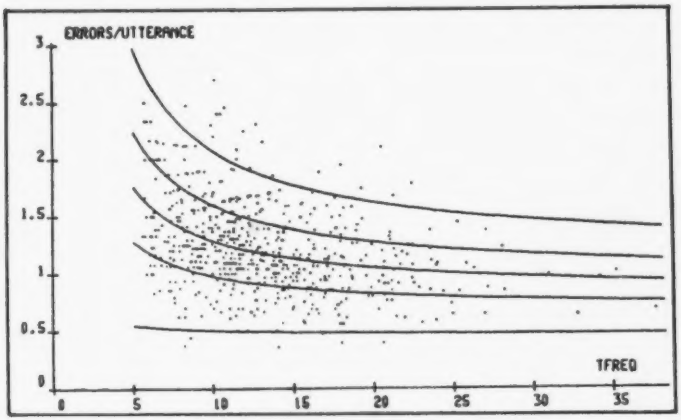
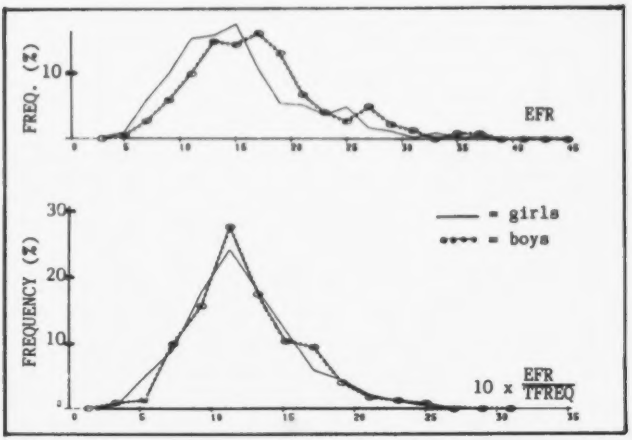


Figure 7

Distributions of the number of errors(EFR) and EFR/TFREQ.



Bibliography

Folland, D., and D. Robertson 1976. "Towards objectivity in group oral testing". English Language Teaching Journal 30: 156-167.

Johnson, N.L. and S. Kotz 1969. Discrete Distributions. New York: John Wiley & Sons.

- 1970. Continuous Univariate Distributions - 1. New York: John Wiley & Sons.

Liski, E.P. 1984. "Regression analysis with two mixed forms of heteroscedasticity". Communications in Statistics - Theory and Methods 13: 1015-1030.

Liski, E.P. and S. Puntanen 1978. "A statistical analysis of the results of group conversation tests in spoken English". Report A 25, Department of Mathematical Sciences, University of Tampere, Finland.

- 1985. "A percentile regression model with an application to error frequency in group conversation tests". The Canadian Journal of Statistics 13, No. 1, in press.

- 1983. "A study of the statistical foundations of group conversation tests in spoken English". Language Learning 33: 225-246.

- 1984. "Statistical modelling of the group conversation tests in spoken English". An abstract in the Proceedings of the 7th World Congress of Applied Linguistics, Brussels, Belgium, August 4 - 10, 1984.

Rao, C.R. 1973. Linear Statistical Inference and Its Applications. Second Edition. New York: John Wiley & Sons.

THE DEVELOPMENT OF READING COMPREHENSION IN FL TEACHING - COMBINING KNOWLEDGE AND SKILL

Madeline Lutjeharms (Vrije Universiteit Brussel)

1 Introductory remarks

After a period of stressing communication and social relations in FL teaching, which meant stressing oral proficiency, reading comprehension has again come to the fore. Certainly, this is in some part due to the frustrations arising from the communicative approach, which leads to such phenomena as pidginization and fossilization as learners are linguistically overstrained, or to oversimplified forms of communication. In productive skills about 85% of errors seems to depend on the level of FL proficiency (Esser 1984 : 151f.). In reading, this percentage might be far lower, as other factors like reading comprehension in L1 or extratextual knowledge will influence the level of comprehension. The fact that communication is two-sided, i.e. that understanding also plays a part in it, has often been overlooked. Yet another reason why the receptive skills have come to the fore might be the socioeconomic situation : the fact of being better informed could be important in finding a job. New trends in research precede the development of new methods. The already extensive literature on reading has increased enormously lately (Pugh 1982 : 7 f.).

Reading and listening are seen not only as goals in themselves, but in the "receptive method" also as the effective first stage of language acquisition and learning in general (Schouten-van Parreren 1983 : 22 ff.). This method claims that exposure to the language helps to develop cognitive models of lexical use and grammatical structures - what we often call "linguistic feeling". That reading and listening are receptive and active skills is a conception now largely accepted. Learners can react to the information, if necessary in L1. Use of L1 is another new and useful trend in FL teaching.

2 Reading while listening

A highly discussed subject is the question of whether students in a reading comprehension course should listen to the text being read aloud while they themselves are reading it. There are arguments against this. Listening while reading could contribute to the wrong impression that a text has always to be read linearly and completely. Neither does it improve reading speed. On the other hand, silent reading is accompanied by speech muscle activity. "Whether this trace of subvocal activity or inner speech ever completely disappears is highly controversial" (Pugh 1978 : 25, cf. 25 ff.; Bennet 1975 : 45; Desselmann et al. 1981 : 259). The more unfamiliar the language, and, one might add, the word, the more throat activity there is.

This is useful, as it supports the processing of information. Reading while listening develops fluency in foreign languages and might improve reading comprehension (Leong 1984 : 43; Pugh 1978 : 27 f.; Schouten-van Parreren 1983 : 26; cf. Leont'ev 1974 : 87, 103). For memorization it is helpful to have more than one channel for the information; another argument for combining reading and listening (Desselmann et al. 1981 : 149). This does not mean that every text should be read aloud or always the whole text, but - especially at the beginning of a reading course - texts, or parts of texts, can be read aloud often, though in general only after a first quick silent reading by the learners. However, it has been argued that it might be better to restrict the reading aloud to short or easy words in the beginning (Nehr 1984 : 82 ff.). Didactic arguments for doing some listening while reading are that a student in a reading course might want to acquire other skills later and that variety in method is a strong motivational factor in itself.

3 What is reading?

What reading comprehension consists of is not very clear. The basic mental processes have not been explained. The problem with reading research is that the reading process cannot be studied directly. There is a lot of research, which started mainly in Europe in the last century but the bulk of which has been done in the USA in this century. Nearly all the research is about reading comprehension in L1 and much of it does not go beyond the sentence level. With interest in text linguistics this has now changed. The difficulties in constructing theories about reading show that we have not yet come very far. This difficulty is perhaps illustrated by the fact that introspection, which was used by Ebbinghaus for memorization with himself as a subject in the last century (Kintsch 1970 : 139 f.) is now being used again with several subjects informants. Introspection can be a teaching device in class too, if one student explains to another one how he or she arrives at comprehension of a certain passage. An interesting survey of research into reading is given by R. de Beaugrande (1981 : 261 ff.). He approaches the subject from the cognitive point of view, stressing information processing and text structure. He leaves out the motor aspects of eye movement and subvocalization. These latter aspects are not important for FL learning, of course, as long as we are concerned with languages with the same kind of writing. The psycholinguistic approach partly overlaps the cognitive approach, but pays more attention to syntactic and semantic aspects, as well as to memory and the role of the senses. In the communicative approach the social side of reading is stressed, with interest being shown in the reading goal, in motivation and in the reader's expectations and the author's intentions. All approaches can be helpful for teachers.

4 Reading a foreign language

Since we can't study them separately the question of whether there are subskills in reading and, if so, what they are has not been

resolved (e.g. Pugh 1978 : 75 ff). For didactic purposes it might be sensible to presume that there are subskills, for "whenever there is some kind of structure in the learning material, verbal learning is facilitated" (Kintsch 1970 : 427). In FL teaching we are normally dealing with learners who have already achieved a degree of proficiency in the reading skill itself. This means that a number of subskills can be adopted for reading a FL. "The native language assumes the role of a meta-model of language itself, i.e. of a set of expectations about how any language is organized" (de Beaugrande 1984 : 12). We can apply cognitive structures or reference frames like scripts or schemes - abstract concepts that give a model of the world - when reading in the FL. Processes of scientific thinking, for example, are more or less independent of the linguistic system and "such textual cues as paragraph organization, definition and classification are similar in different languages" (Ulijn 1984 : 72). We know what a sentence or a word is. The reader takes his background of experience with him, his "knowledge of the world". Although there can be differences in the culture of the FL, these are only minor compared to the linguistic problems. But more than in L1, reading a FL is a problem-solving process, the construction of hypotheses is more important, when linguistic factors are not fully under control. When reading a FL it is necessary to be able to use all the information provided by the text in order to compensate for the linguistic gaps. This means that guessing is a very important skill. Affective factors may play a different role in reading the FL than in L1. The interest in the FL or fear of the difficulty of the task will influence the acquisition of the skill.

It is evident that it is in the language-based part of the reading process that the new factors are to be found. There it is probably better to talk of knowledge than of skills, although the two cannot be separated completely. New words and phrases have to be recognized, which implies knowledge of lexis, morphology and syntax. But a large part of this knowledge will have to function automatically, i.e. on the unconscious level. The difficulty lies mainly in the surface text. Text structure will be less of a problem, although differences in culture, e.g. in scientific tradition, do play a role.

5 The role of guessing

The main skill for a rational development of reading comprehension in a FL is probably guessing. When teaching German to Dutch-speaking students one starts to train guessing because it is so evident, the languages being very close. Learners will understand a number of words without guessing, but when guessing - normally rewarded with a high rate of success - the vocabulary is enlarged considerably with the help of L1. The training of contextual guessing where L1 doesn't help is then the next step. Van Parreren and Schouten-van Parreren call this "the most specific skill for foreign language reading" (1981 : 236). This means that the use of as much information as context - and former experience - can provide, has to be trained, as well as analyzing word forms to use all the information contained in parts of the word. Moreover, the use of guessing is a

good strategy for vocabulary learning, as mental activity on the part of the learner in finding the meaning will improve retention. Help from the teacher may be necessary though, since it is very important to guess the right meaning, in order to prevent error memorization (Schouten-van Parreren, and van Parreren 1979 : 265 f.). Van Parreren and Schouten-van Parreren give a survey of where guessing can go wrong. They give a hierarchy of levels for the guessing : 1. syntactic, 2. semantic (context and knowledge of the world), 3. lexical (word form, e.g. affixes or cognates), 4. stylistic (1981 : 238 ff.). This hierarchy is useful for the teacher, who will have to start his questioning from a given level according to the nature of the difficulty encountered. Sometimes, but not always, all levels may be necessary. In this case the teacher could start with the question "what kind of word is it?" before going on to "what is the text, paragraph or sentence about?"

Apart from a number of special vocabulary exercises, we work with a lot of authentic reading material. Most frequent words will be learned incidentally, because they occur so often. The students are asked to mark difficult words clearly in the text, so that revision is always in context. Difficult words include not only deceptive cognates, but also those where contrast is lacking because they look familiar through their resemblance to other words or to parts of other words, either in L1 or in the FL (Juhász 1970 : 92 f. and Lutjeharms 1981 : 88). Guessing skills, of course, are not only important for vocabulary. The reader needs them on the text level as well. But guessing on both levels interacts. A wrong hypothesis about text content leads easily to wrong decoding of vocabulary, and an error in word meaning often leads to a complete misunderstanding of a passage or sometimes of a whole text, as the reader tries to find coherence in the text. Part of the work of the teacher lies in making the learner conscious of guessing skills, since the learner will not always apply them automatically to the FL situation. Besides, they will enhance the preparedness to start reading a FL.

6 Contrastive method

The use of a contrastive method will in general improve reading comprehension. It means that the learners are made conscious of all they already know before having started to learn the FL, but it also, of course, involves an insistence on the special difficulties of the target group. Positive transfer may be rather limited, when comparing for example Finnish and German, but another FL might be helpful, in this case some experience with Swedish and English. The possibilities of transfer will also depend on the text type; in LSP lexical transfer will normally be more important than in common language.

7 Exercises

If the teaching of reading is part of a course in overall language proficiency, the approach will be quite different from a

course where reading is the only goal. In the case of a strict reading course, some parts of grammar for instance will be left out, whereas others will be treated differently. Before proceeding to the exercises it should be mentioned that we are here concerned not with all text-based exercises, but only with those intended to develop FL reading comprehension. These can be classified as : 1. exercises preparing and accompanying the reading, 2. exercises on global text comprehension or extensive reading, 3. exercises on detailed text comprehension or intensive reading. This is a kind of hierarchy, as one level will support the next one, but it does not always mean that the exercises must be in a rigid order. Variety is important. Obviously, not all the exercises to be presented have been devised by the author. The reader will certainly know most of them already and be able to add other types. The intention is to show what kind of exercises proved to be useful in our experience, which consists mainly of teaching LSP to university students. Obviously, the main reasons for adopting certain methods or materials should be : firstly, the needs and motivation of the learners and, secondly, the motivation of the teacher and his/her belief in the method.

Before starting to work on any exercise or assignment, the learners should get clear instructions about the objectives and procedures. Insight into reading styles can be a help (on reading styles see Pugh 1978 : 52 ff.). It is important for them to know what they are doing and why. The learning attitude and the level at which information is processed will change according to the assignment.

7.1 Exercises preparing and accompanying the reading

The bulk of these are vocabulary exercises. There are strong indications that most of the difficulties experienced in reading a FL are lexical rather than syntactical, which does not however mean, especially at the intermediate level, that syntax is no problem (Ulijn 1984 : 69 ff.; Bhatia 1984 : 91 f.). In reading L1, word knowledge appears to be a good predictor of reading comprehension (Thorndike quoted by Pugh 1978 : 77), just as word recognition for the FL (Deering 1982 : 182 ff.). Exercises that ask learners to manipulate words and to reflect on their structure and/or meaning can be expected to improve reading skills, since doing something with a word results in more learning success than just trying to memorize it.

7.1.1 Vocabulary exercises

7.1.1.1 The first vocabulary exercise is meant as an introduction to the use of a dictionary. Words are given in declined and conjugated forms, or as derivations and compounds that are not in a dictionary in this form, verbs with separated prefixes etc. The dictionary form has to be found, often by guessing, but indications for succesful guessing are given, e.g. how to find the infinitive of an irregular verb. Dictionaries are used in class from time to time, and for a limited time at the examination. When students ask for the meaning of a word while working with a text, they are asked to give the dictionary-form of it and to identify the kind of word it is, e.g. a verb,

on the basis of form and/or position in the sentence.
7.1.1.2 For frequent difficult words and deceptive cognates transformation and recognition exercises and exercises filling in the gaps have been made. All examples are extracts from longer exercises.
- Transformation, e.g. for the modal verbs "sollen/müssen" ("sollen" is a deceptive cognate).
  - Die Änderung hat das Ziel das Stadtverkehrssystem zu verbessern. Die Änderung soll/muss ....
  - Es ist unbedingt notwendig, in Zukunft eine Überforderung zu vermeiden. Man soll/muss
- Recognition : which words or expressions have the same meaning?
  - modal verbs

	paraphrases
1. Ich muss diesen Auftrag durchführen	a. Ich habe die Absicht
2. Ich darf diesen Auftrag durchführen	b. Man erwartet es von mir
etc.	etc.

  - internationalisms and difficult German words or deceptive cognates

a. Thema	1. Verfahren
b. Faktum	2. Vorgang
c. Prozess	3. Tatsache
d. Methode	4. Gegenstand

  The puzzle aspect of these exercises makes them very popular.
- Filling in the gaps
  First the meaning of a group of words is explained, then a text with gaps for these words follows. These texts or sentences are usually not authentic, as they would be difficult to find. There are for instance exercises on the word group "öffentlich, offen, offenbar", which are all deceptive cognates, on the numerous meanings of "Wirtschaft" and "wirtschaftlich", or on the nouns with the prefix "Er-", where lack of contrast comes into play. Crossword puzzles on difficult vocabulary are another kind of gap-filling exercise and are certainly more popular.

7.1.1.3 Language awareness tasks have been constructed on word formation, to give insight into morphological variants and lexical compounding. They are meant as a kind of very guided guessing. Mastery of the process of nominalization, affixation and compound word formation enlarges vocabulary enormously in an efficient way. Such exercises structure knowledge, thereby reducing the memory load.
- compound word formation

What word is defined by what paraphrase?

1. der Transportbehälter	a. Behälter zum Transportieren
2. der Behältertransport	b. Transport in (von) Behältern
3. der Kontrollraum	c. Kontrolle eines Raumes
4. die Raumkontrolle	d. Raum, in dem kontrolliert wird

- derivation

exercise on the meaning of common affixes

die Herstellung	herstellen	herstellbar	die Herstellbarkeit
die Verwendung	.....	.....	.....

exercise on adjectives

die Härte	härten		.....
die Wärme	(er)...		.....
die Grösse	vergrössern		.....
die Breite	.....	(cf. verbreiten)	.....

to introduce adjective affixes : pick out opposite pairs :

lebhaft	leblos, tot, ruhig
lebendig	lebensunfähig
lebensfähig	ruhig

This kind of exercise is used also after a reading assignment with a text. A number of nouns and verbs are underlined in different ways; the nouns have to be transformed into verbs and vice versa. This can be done only after systematic derivation exercises. An easier exercise is to look for all compounds or derivations with a certain item, e.g. "Arbeit" in a strike leaflet. The meaning of these words is then discussed.

7.1.2 Exercises on grammatical points

7.1.2.1 Perception exercises
Complete beginners are asked to look for inflectional indicators with the help of declension tables. In a short text all visible declined forms have to be underlined. A discussion follows on what e.g. an —er - ending can mean. This is done also with singular and plural nouns, prepositions etc. Of course, here texts are needed with frequent and varied appearances of these endings or words.
7.1.2.2. Transformation exercises on text segments, that are used mainly for reinforcement.
- questions that are asked in order to change the case

Text : "Das Geheimnis[1] des ungebrochenen Schweizer Wirtschaftserfolges, darin sind sich[2] Arbeitgeber und Gewerkschaften weitgehend einig, liegt in der[3] erprobten Kompromissbereitschaft der Tarifparteien". (Die Zeit).

1. Wozu führte die Kompromissbereitschaft der Tarifparteien?
2. Mit wem sind sich die Gewerkschaften weitgehend einig? (Antwort mit dem bestimmten Artikel)
3. Was gibt es zwischen den Tarifparteien? (Antwort mit dem unbestimmten Artikel)

A quick recognition of inflectional indicators is the learning goal for these exercises. Although many inflectional indicators are redundant, they may be important in some cases, esp. in social science texts, where a sentence starting with an object is less rare than in natural science texts. These exercises clearly involve productive skills, but this can be stimulating and it supports memorization and automatization. Another grammatical exercise consisting of questions on a text is constructed in such a way that all the answers have to be subordinate clauses.
- A different type of exercise consists of transformation or construction of nominalizations. The construction is at the same time an exercise on the genitive attribute.

Transformation of nominalizations

Erforschung und Training des Zuhörens.
⟶ Er hat das Zuhören erforscht und trainiert.
Die Aufnahme von Gehörtem. ⟶ Man ....

Construction of nominalizations

Zahllose Termine werden revidiert.
⟶ Die Revision/Revidierung zahlloser Termine.
Er hielt das erste "Lauscher-Trainingsseminar ab. (Die Presse) ⟶ .....

This can also be done without a question sheet, just by underlining in the text.

7.2 Reading exercises and assignments

A real reading exercise consists of work on the information provided by a text. If L1 and the target language have little in common, it might be necessary to start with very easy, short texts, like easy book titles, lists etc. with lots of cognates, perhaps with translations to match or with instructions to underline words that are recognized.

7.2.1 Global text comprehension or extensive reading

7.2.1.1 To improve comprehension and information recall it can be useful to start with an "advance organizer" (term from Ausubel), consisting of "appropriate, relevant and inclusive introductory material which provides anchorage for the integration and retention of new, more differentiated information" (Boekaerts 1980 : 131). These can be pictures, statistics or questions by the teacher. If the subject matter belongs to an area the learners are familiar with, the advance organizer may be reduced to a short introduction or introductory remark to concentrate the attention on the subject. A second step is to make use of the ability to predict, to make conjectures on the possible contents of the text. This ability can be taught and improved by guided questioning. It is certainly an essential factor in effective reading (Friedman and Rowls 1980 : 6 ff.). You can start from the title (in newspaper articles often rather from the subtitle), from the introductory paragraph or with a quick skimming. The importance of starting to work with a text as a whole and proceeding from this global view has to be stressed. This is often called the top-down procedure. The predictions have to be very guided in the beginning, as reading can be disturbed by false expectations. It may be necessary to give the meaning of some important words that can't be guessed successfully in advance, in order not to overstrain the learners. On the other hand they have to be made conscious of all they know already, before reading the text, on the basis of frames of reference and former experience.

The first reading task can be to skim the text for an overall impression. You can also combine reading styles, e.g. by starting from a given reading goal, like scanning for a word or a name or search reading to locate information, and switching over to closer reading as soon as there is a passage where the word occurs or the relevant information might be found. It is often useful to set time limits for these tasks. To test text comprehension it is necessary to use L1 with beginners in the FL, as we don't want to use productive skills. The only ways to avoid this would be pictures, figures or a diagram, or perhaps multiple choice questions in the FL, but in this last case we have a completely different and far more difficult reading situation, because of lack of context. Besides, to test comprehension with the help of L1 is probably the most precise way of doing it. Another reason for using L1 is that in long term memory we often store the information without the mediating linguistic material of the FL.

7.2.1.2 The classic, very general questions on text contents are useful from time to time. The answers may be in L1, especially for beginners.
7.2.1.3 A good exercise for global text comprehension is to identify the different parts of the text structure. This exercise can be guided by giving statements on these parts that have to be related to the text or put into the right order. Use of top-level structure provided in the text improves comprehension skills (Meyer, Brandt, and Bluth 1980 : 96 ff.).

7.2.2 Exercises on detailed text comprehension or intensive reading.

After an extensive reading assignment the students can be asked to read the text rather quickly and mark where they don't understand, i.e. words or structures that interfere with comprehension on the sentence level. They have to learn that they don't always have to understand each word. On the other hand they should be trained in discovering what they don't understand without help. When working with long texts, exercises on intensive reading can be made for parts of the text only. The reading styles for detailed text comprehension will be receptive and responsive; the assignments are partly post-reading exercises. There are many possibilities, which are often inspired by the text itself, by its structure, contents, style or vocabulary. Some examples follow.
7.2.2.1 Ask the learners to distinguish main points from subsidiary points by underlining all important and new information and leaving out details (often examples or explanations), function words or phrases, redundancies etc. This is a good way to learn summarizing. A variation on this is to search for words that label important or essential information, if you find a suitable text for the purpose.
7.2.2.2 Vocabulary questions to point out the importance of the context in finding deceptive cognates are useful. Internal consistency of the text is the only way to discover them. To prevent frequent use of the dictionary, guessing with the help of both L1 (when it is sensible) and context is trained.
7.2.2.3 Complex sentence structures are analyzed, simply starting from the main clause and from there reconstructing the sentence. Yet syntax is not the main problem. There are only a few difficult points, which are discussed, each time they appear. For this purpose it is useful to number the lines, before copying a text.
7.2.2.4 Some texts are suited to the task of looking for certain forms, words or contents, such as those expressing uncertainty or a certain point of view of the author. Other possibilities, especially for beginners, include looking for negations or statements of place or quantities etc. These exercises are mainly meant to have the learners reread the text several times with different assignments for fixation of vocabulary or structures. A variation on this is to give the learners paraphrases of words or sentences, for which they have to find the equivalents in the text.
7.2.2.5 Questions on the contents of the text can also be used for intensive reading. The answers can once again be either in L1 or the FL. Questions can be asked to reproduce text contents, on the intentions of the author, or to compare, to give personal reactions, to classify the information in a diagram ec. On manipulative or dema-

gogic texts, such as those in "Bild"-Zeitung, revealing questions are made to guide the students in discovering the manipulation by inference. Sometimes students have to ask the questions themselves.
7.2.2.6 A variant on the questions on content matter are the statements on the text, whereby the learners have to decide whether they are in line with the text, different from it or extraneous to it. You can ask the learners to mark the respective passages in the text.
7.2.2.7 Sentences from the text or from a summary of the text can be mixed up. The learners have to find the original order with the help of the text.
7.2.2.8 Specific exercises filling in missing words, e.g. on signal words, like conjunctions, are difficult to make, as it is not easy to find adequate texts. A solution to this problem is to make questions on the text and provide the answer as well, leaving a gap for the conjunction. A list of necessary conjunctions is given beforehand. Other exercises on text cohesion include searching for all the ways in which the subject of the text, for example, is expressed (the noun or name itself, paraphrases, pronouns etc.) or to find the antecedents of pronouns. An interesting exercise, devised by Köhler (1980 : 27), is a text in which the syntax has been changed so that the continuity has been partly lost. The text has to be made coherent again.

Note

I should like to thank my colleague Frank Winter for reading through the English text.

Bibliography

Bennet, W.A. 1975. "Psycholinguistics and the teaching of reading". IRAL 13 : 43 - 50.

Beaugrande, R. de. 1981. "Design criteria for process models of reading". Reading Research Quarterly 16 : 261 - 315.
- 1984. "Reading skill for foreign languages : a processing approach". In Pugh, A.K., and J.M. Ulijn (eds) 1984. 4 - 26.

Bhatia, V.K. 1984. "Syntactic discontinuity in legislative writing and its implications for academic legal purposes". In Pugh, A.K., and J.M. Ulijn (eds). 1984. 90 - 96.

Boekaerts, M. 1980. "Analysis of the learning process : what can the teacher learn from such an analysis?" In Eichheim, H., and A. Maley (eds). 1980. Fremdsprachenunterricht im Spannungsfeld zwischen Gesellschaft, Schule und Wissenschaften. Pariser Werkstattgespräch 1979. München, Goethe-Institut. 116 - 145.

Deering, J. 1982. "Leessnelheid in de moedertaal (Nederlands) en in een vreemde taal (Frans)". Toegepaste Taalwetenschap in Artikelen 13 : 182 - 201.

Desselmann, G., and H. Hellmich (Autorenkollektiv unter Leitung von). 1981. Didaktik des Fremdsprachenunterrichts. Deutsch als Fremdsprache. Leipzig : VEB Verlag Enzyklopädie.

Esser, U. 1984. "Fremdsprachenpsychologische Betrachtungen zur Fehlerproblematik im Fremdsprachenunterricht". Deutsch als Fremdsprache 21 : 151 - 159.

Friedman, M.I., and M.D. Rowls. 1980. Teaching Reading and Thinking Skills. New York, etc. : Longman.

Juhász, J. 1970. Probleme der Interferenz. Budapest, Akadémiai Kiado.

Kintsch, W. 1970. Learning, Memory and Conceptual Processes. New York etc., John Wiley & Sons, Inc.

Köhler, C. 1980. "Wortschatz aus Industrie und Technik (1)". Sprachpraxis 17,5 : 26 - 27.

Leong, C.K. 1984. "Coding processes in reading alphabetic and morphemic scripts". In Pugh, A.K. and J.M. Ulijn (eds). 1984 : 35 - 47.

Leont'ev, A.A. 1974. Psycholinguistik und Sprachwissenschaft. Stuttgart etc., Kohlhammer.

Lutjeharms, M. 1981. "Zum Testen des fachorientierten Leseverständnisses : Ein Beispiel." In Culhane, T., Klein-Braley, C., and D.K. Stevenson (eds). 1981. Practice and problems in language testing. University of Essex, Dept. of Language and Linguistics. 88 - 94.

Meyer, B.J.F., Brandt, D.M., and G.J. Bluth. 1980. "Use of top-level structure in text. Key for reading comprehension of ninth-grade students. Reading Research Quarterly 16 : 72 - 103.

Nehr, M. 1984. "Audiolingual behaviour in learning to read foreign languages : reports on laboratory experiments and a pedagogical study". In Pugh, A.K., and J.M. Ulijn (eds). 1984. 82 - 89.

Parreren, C.F. van, and M.C. Schouten-van Parreren. 1981. "Contextual guessing : a trainable reader strategy". System 9 : 235 - 241.

Pugh, A.K. 1978. Silent reading. An introduction to its study and teaching. London : Heinemann.

- 1982. "Recent trends in reading research and teaching". Toegepaste Taalwetenschap in Artikelen 13 : 7 - 29.

- and J.M. Ulijn (eds). 1984. Reading for professional purposes. London etc. : Heinemann.

Schouten-van Parreren, M.C. and C.F. van Parreren. 1979. "De verwerving van een vreemdtalige woordenschat. Een literatuurstudie". Levende Talen 341 : 259 - 269.

- 1983. "Wisseling van de wacht in de vreemdetalendidactiek?". Levende Talen 378 : 22 - 29.

Ulijn, J.M. 1984. "Reading for professional purposes : psycholinguistic evidence in a cross-linguistic perspective". In Pugh, A.K., and J.M. Ulijn (eds). 1984. 66 - 81.

EVIDENCE FOR UNIQUE ARTICULATORY AND AUDITORY DIMENSIONS IN ESL DATA

Kyle Perkins, Sheila R. Brutten, and Paul J. Angelis (Southern Illinois University)

1 Introduction

In language testing, unidimensionality refers to the assumption that subjects' responses to test items are a function of one and only one construct measured by the test under consideration. On the other hand, multidimensionality refers to the assumption that item responses are a function of more than one construct. The notion of dimensionality has been extended in recent years to include investigations of the construct validity of language tests and the hypothesis of a single language proficiency dimension or factor underlying different language tests.

Several investigators have employed principal components analysis in their analysis of test data and have reported their findings in support of the single unitary factor hypothesis--also known as the unitary competence hypothesis (e.g., Oller 1979; Oller and Hinofotis 1980). On the other hand, Bachman and Palmer (1981) refute the unitary factor hypothesis, instead presenting evidence for a model with two correlated trait factors and three method factors. In addition, Upshur and Homburg (1983) present evidence in favor of multiple-factor models of language proficiency.

The purpose of this investigation is to analyze a data set to determine whether evidence can be found to support or to refute the unitary factor hypothesis of language proficiency.

2 Methodology

The subjects for this investigation were 50 adult students enrolled in the most advanced full-time level of instruction at the Center for English as a Second Language at Southern Illinois University at Carbondale. Native languages represented in the subject pool included Spanish, Arabic, Chinese, Japanese, Korean, Malaysian, Thai, Greek, Tamil, Polish, French and Yoruba.

3 Instrumentation.

Data were elicited in the language laboratory with the following instruments:

The Wepman Auditory Discrimination Test consists of 40 word pairs equated for length; 30 of the pairs differ in a single phoneme, while 10 pairs are identical. The subject must indicate whether the spoken pair is the same or different.

Three tests from the Seashore Measure of Musical Talents were employed. In the Seashore test of pitch, the subject is presented 50

pairs of tones, and the subject must determine whether the second tone is higher or lower in pitch than the first. The tones are at about 500 cycles and have a duration of 0.6 second each. Frequency differences in cycles between the tones in the pairs range from 2 to 17. In the Seashore test of loudness, the subject is presented 50 pairs of tones, and the subject must indicate for each pair whether the second tone is stronger or weaker than the first; the frequency is held constant at 440 cycles. The intensity differences in decibels between the tones in the pairs range from 0.5 to 4.0. In the Seashore test of rhythm, the subject is required to indicate whether a series of 30 pairs of rhythmic beats are the same or different.

The sentence repetition test involves memory; it is a measure of the immediate recall of sentence length verbal material. The test consists of 26 tape-recorded sentences of increasing length and complexity from "Mary agreed" to "the sun never reaches the ground through the overhead canopy of trees and vines." The subject must repeat each sentence immediately after two presentations. There is a twenty second interval between the second presentation and the next sentence. For scoring, one point was allowed for each sentence. To receive credit, the subject must repeat each word in the sentence; the words must be correctly sequenced, and all bound morphology must be present. For example, if a subject repeated "Mary agree" for the stimulus "Mary agreed," the subject receives no point for that sentence. The 26 sentences contained a total of 216 words. The longer sentences were included in the test due to Wanner's (1974) observation that listeners lose the exact wording of a sentence when only 16 syllables (for native speakers) intervened between the presentation of a sentence and its test. Pilot work for this research indicated that 7 to 8 syllables is the threshold for ESL subjects.

The Test of Spoken English (TSE) was used to evaluate the oral English speaking proficiency of the subject pool, and it consists of seven sections, each of which requires a different speaking activity. The following description is found in the Guide to SPEAK (1982:6-7):

> The first section is an unscored "warm-up" in which the examinee responds orally to a few brief biographical questions provided on the test tape.
>
> In the second section, the examinee is allowed time for preliminary silent reading of a passage of about 125 words and then is instructed to read the passage aloud. Scoring is based on pronunciation and overall clarity of speech.
>
> In the third section, the examinee is asked to complete a series of ten partial sentences in a way that conveys meaning and is grammatically correct.
>
> The fourth section of the test consists of six line drawings that tell a continuous story. After studying the drawings briefly, the examinee is asked to tell the story that is depicted, using past tense narration.
>
> In the fifth section, the examinee looks at a single line drawing and answers several spoken questions about the picture.
>
> The sixth section consists of a series of spoken questions intended to elicit relatively free and somewhat more lengthy

responses. Questions requiring both straightforward descriptions of common objects and fairly open-ended expressions of opinion are included. The linguistic quality and adequacy of communication, not the factual content of the responses, are at issue in scoring.

In the seventh and final section, the examinee sees a printed schedule, such as the outline for a course or a conference, and is asked to describe the schedule aloud, as though informing a group of listeners.

The TSE rating schedules are presented in Appendix A.

These instruments were chosen because previous research has shown that auditory factors and memory span are highly related to attained second and foreign language proficiency (e.g., Guilford and Christensen 1956; Carroll 1941; Holzinger and Harmon 1938; Karlin 1942; Carroll 1962; Pilmsleur, Stockwell and Comrey 1962; Leutenegger and Mueller 1964; Leutenegger, Mueller and Wershow 1965; Blickenstaff 1963; Suter 1976; Purcell and Suter 1981).

4 Analyses and Results

The means, standard deviations, and estimates of reliability for all measures are presented in Table 1. The correlation matrix is displayed in Table 2.

To determine whether one or more factors were manifested in the test data, we selected the principal factor solution with iterations. The criterion for the number of factors to be retained was an eigenvalue for each factor equal to or greater than 1. The principal factor solution with iterations is presented in Table 3.

We selected the varimax rotation option to make each factor define in a unique manner a distinct cluster of inter-correlated variables in the factor structure, provided that such clusters do indeed exist. If the unitary factor hypothesis is to be confirmed by this research, then only one such cluster should be found in the factor structure. If more than one distinct cluster of intercorrelated variables exist, this research will disconfirm the unitary competence hypothesis.

We wanted to be objective about how large, significant or salient a factor loading must be in the varimax rotated solution to merit attention. Following Gorsuch (1974), we established that with samples as small as ours, a factor loading of $\pm$ .64 is significant. Gorsuch suggests that a rough check on the minimum salient loading is provided by doubling the appropriate standard error. For our sample, the minimum significant correlation coefficient ($p<.01$) with an n of 50 is .32; therefore, loadings greater than .64 are considered salient or significant.

The varimax rotated factor matrix is presented as Table 4. Two factors were retained and rotated. All the speaking measures save for sentence repetition load significantly on Factor 1 which we will

refer to as a speaking or articulatory factor. We assume that the reason why the sentence repetition test did not significantly load on the articulatory factor is that it is also a measure of listening comprehension and short-term memory. We shall refer to Factor 2 as the auditory discrimination factor because there were moderate loadings by the different measures of auditory discrimination. Only one measure--the Seashore Loudness test--loaded significantly but by a hair-thin margin. The results of Table 4 seem to indicate that the unitary competence hypothesis was disconfirmed by this research. If there were support for the unitary competence hypothesis, we should have seen both the auditory and articulatory measures load significantly on one factor.

We must point out that these results should be interpreted with caution for the following reasons: (1) we believe that one should have at least 30 to 35 subjects per measure before submitting a data set to factor analytic solutions. In our case that would mean at least 270 subjects (we had 50): (2) there are doubtlessly more cautious and conservative methods of determining how high a factor loading must be to be significant, that is, we may have set the criterion factor loading too low.

For the reasons set out in the preceding paragraph we conducted two additional non-factor-analytic analyses. First, we corrected the correlation matrix for attenuation (Table 5). Then we partitioned the correlation matrix into three sections: (1) a cluster of intercorrelations of the listening discrimination measures---what we term a monotrait-heteromethod cluster. All the measures in this group purport to measure listening discrimination but employ different methods such as differences of loudness, pitch, phonemes, etc.; (2) a cluster of intercorrelations of the speaking or articulatory measures--a monotrait-heteromethod cluster; and (3) a cluster of intercorrelations of both the auditory and articulatory measures--a heterotrait-heteromethod cluster. The correlation coefficients for each cluster were averaged. As Table 5 indicates, the mean correlation for the articulatory measures (.723) is over twice as large as the mean for the auditory discrimination measures (.340) and almost seven times greater than the mean for the heterotrait-heteromethod cluster (.128). We believe that the data in Table 5 corroborate our factor-analytic results. If there were only one factor underlying the data, we would expect that the mean correlation for the heterotrait-heteromethod cluster would have been more in line with the other two means.

We did not submit these means to any statistical test of significance, because we are not sure if such a test would be appropriate. However, if such a test were deemed appropriate, we recommend that the correlation coefficients be transformed to a normal interval scale by using the Fisher Z transformation before they are averaged.

A further non-factor-analytic procedure involved determining the proportion of the reliable variance which is shared between pairs of variables. The procedure involves squaring the correlation "corrected" for unreliability:

$$r^2 = \left[\frac{r_{xy}}{\sqrt{r_{xx}\, r_{yy}}} \right]^2$$

where r_{xy} is the correlation coefficient obtained between the x and y variables

r_{xx} is the estimate of reliability of the x variable and

r_{yy} is the estimate of reliability of the y variable.

The proportions of shared reliable variance are reported in Table 6. We believe that the data displayed in Table 6 also corroborate the factor analysis results and interpretations. If there were a single unitary factor underlying these data, we would expect the shared reliable variance between auditory and articulatory measures to be as high or nearly as high as the variance between any two auditory variables or any two articulatory variables. Table 6 indicates that this is not the case.

A related issue--although a thorough treatment would go beyond the scope of this paper--is the amount of variance which is attributable to the trait and the amount of variance which is attributable to the test method. Bachman and Palmer's (1981) research indicates that method effects can account for substantial amounts of variance. The question of whether the trait/method variance proportion is 50/50, 67/33, 25/75, etc., remains open.

In summary, the results of this investigation suggest that there are two distinct factors present in the data analyzed--an auditory factor and an articulatory factor. With more data elicitation and more sophisticated analysis such as multitrait-multimethod validation and confirmatory factor analysis, one might find evidence for a method factor. The present analyses and interpretations tend to disconfirm the unitary competence hypothesis proposed by Oller (1979) and Oller and Hinofotis (1980) and seem to be more in accordance with the multiple-factor models proposed by Upshur and Homburg (1983) and Bachman and Palmer (1981).

APPENDIX A

TSE Rating Schedules

Pronunciation

0 - Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be unintelligible

1 - Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be occasionally unintelligible

2 - Some consistent phonemic errors and foreign stress and intonation patters but speaker is intelligible

3 - Occasional nonnative pronunciation errors but speaker is always intelligible

Grammar

0 - Virtually no grammatical or syntactical control except in simple stock phrases

1 - Some control of basic grammatical constructions but with major and/ or repeated errors that interfere with intelligibility

2 - Generally good control in all constructions with grammatical errors that do not interfere with overall intelligibility

3 - Sporadic minor grammatical errors that could be made inadvertently by native speakers

Fluency

0 - Speech is so halting and fragmentary or has such a nonnative flow that intelligibility is virtually impossible

1 - Numerous nonnative pauses and/or a nonnative flow that interferes with intelligibility

2 - Some nonnative pauses but with a more nearly native flow so that the pauses do not interfere with intelligibility

3 - Speech is as smooth and as effortless as that of a native speaker

Comprehensibility

0 - Overall comprehensibility too low in even the simplest type of speech

1 - Generally not comprehensible due to frequent pauses and/or rephrasing, pronunciation errors, limited grasp of vocabulary, or lack of grammatical control

2 - Comprehensible with errors in pronunciation, grammar, choice or vocabulary items, or infrequent pauses or rephrasing

3 - Completely comprehensible in normal speech with occasional grammatical or pronunciation errors

Table 1

Means, standard deviations, and estimates of reliability

Test	Number of subjects	Number of items/points	Mean	SD	Type of reliability	Estimate of reliability
Wepman Auditory Discrimination	50	40	36.54	1.876	KR-20	.320
Seashore Pitch	50	50	31.02	7.844	KR-20	.867
Seashore Loudness	50	50	41.26	9.167	KR-20	.938
Seashore Rhythm	50	30	24.46	4.234	KR-20	.803
Sentence Repetition	50	26	9.64	3.768	Interrater	.910
TSE Pronunciation	50	3	1.63	.468	Interrater	.935
TSE Grammar	50	3	1.65	.522	Interrater	.847
TSE Fluency	50	3	1.70	.492	Interrater	.727
TSE Comprehensibility	50	300	171.20	50.485	Interrater	.826

Table 2

Correlation matrix

	1	2	3	4	5	6	7	8	9
1. Wepman Auditory Discrimination	1.00	.15 p=.144	.24 p=.045	.13 p=.185	.39 p=.003	-.04 p=.380	.05 p=.367	-.08 p=.289	-.05 p=.374
2. Seashore Pitch		1.00	.43 p=.001	.21 p=.073	-.001 p=.497	-.09 p=.264	-.16 p=.126	-.05 p=.372	-.14 p=.166
3. Seashore Loudness			1.00	.29 p=.019	.25 p=.038	.15 p=.150	.11 p=.221	.23 p=.056	.15 p=.150
4. Seashore Rhythm				1.00	.40 p=.002	.33 p=.010	.14 p=.157	.18 p=.109	.27 p=.029
5. Sentence Repetition					1.00	.44 p.001	.38 p=.003	.42 p=.001	.36 p=.005
6. TSE Pronunciation						1.00	.75 p=.000	.77 p=.000	.89 p=.000
7. TSE Grammar							1.00	.62 p=.000	.89 p=.000
8. TSE Fluency								1.00	.74 p=.000
9. TSE Comprehensibility									1.00

Table 3

Principal factor solution with iterations

	Factor 1	Factor 2	Communality	Factor	Eigenvalue	Percent of Variability	Cumulative Percent
Wepman Auditory Discrimination	.05633	.41957	.17922	1	3.51722	74.7	74.7
Seashore Pitch	-.06361	.52877	.28365	2	1.19322	25.3	100.0
Seashore Loudness	.24894	.61374	.43864				
Seashore Rhythm	.33772	.40003	.27408				
Sentence Repetition	.51637	.34792	.38768				
TSE Pronunciation	.92409	-.09647	.86325				
TSE Grammar	.82964	-.17495	.71892				
TSE Fluency	.77699	-.04695	.60592				
TSE Comprehensibility	.96015	.19451	.95972				

Table 4

Varimax rotated factor matrix

	Factor 1	Factor 2
Wepman Auditory Discrimination	-.02372	.42267
Seashore Pitch	-.16208	.50732
Seashore Loudness	.12887	.64964*
Seashore Rhythm	.25632	.45649
Sentence Repetition	.44158	.43896
TSE Pronunciation	.92572*	.07933
TSE Grammar	.84775*	-.01554
TSE Fluency	.77193*	.10026
TSE Comprehensibility	.97960*	-.01016

*significant (salient) factor loading

Table 5

Correlation matrix corrected for attenuation

	1	2	3	4	5	6	7	8	9	
1. Wepman Auditory Discrimination		.285	.438	.256	.723	-.073	.96	-.166	-.097	heterotrait heteromethod $\bar{X}$=.128
2. Seashore Pitch	monotrait-heteromethod $\bar{X}$=.340		.477	.252	.001	-.100	-.187	-.063	-.165	
3. Seashore Loudness				.344	.271	.160	.123	.278	.170	
4. Seashore Rhythm					.468	.381	.169	.236	.332	
5. Sentence Repetition						.477	.433	.517	.415	monotrait-heteromethod $\bar{X}$=.723
6. TSE Pronunciation							.844	.934	.910	
7. TSE Grammar								.789	.956	
8. TSE Fluency									.955	
9. TSE Comprehensibility										

Table 6

Shared reliable variance between variables

	1	2	3	4	5	6	7	8	9	
1. Wepman Auditory Discrimination		.081	.661	.066	.523	.005	.009	.027	.009	heterotrait-heteromethod $\bar{X}$=.071
2. Seashore Pitch			.227	.063	.000	.010	.035	.004	.027	
3. Seashore Loudness	monotrait-heteromethod $\bar{X}$=.279			.578	.073	.026	.015	.077	.029	
4. Seashore Rhythm					.219	.145	.029	.056	.110	
5. Sentence Repetition						.227	.187	.267	.172	
6. TSE Pronunciation							.712	.872	.828	monotrait-heteromethod $\bar{X}$=.572
7. TSE Grammar								.622	.916	
8. TSE Fluency									.912	
9. TSE Comprehensibility										

REFERENCES

Bachman, L. and Palmer, A. 1981. "The construct validation of the FSI oral interview". Language Learning 31: 67-86.

Blickenstaff, C. 1963. "Musical talents and foreign language learning ability." Modern Languge Journal XLVII: 359-363.

Carroll, J. 1941. "A factor analysis of verbal abilities". Psychometrika 6: 279-307.

Carroll, J. 1962. "The prediction of success in intensive foreign language training". In Glaser, R. (ed). Training research and education. Pittsburgh, Pennsylvania: University of Pittsburgh Press.

Educational Testing Service. 1982. Guide to SPEAK. Princeton, New Jersey: Author.

Educational Testing Service. 1982. Test of spoken English. Princeton, New Jersey: Author.

Gorsuch, R. 1974. Factor analysis. Philadelphia, Pennsylvania: W. B. Saunders Company.

Guilford, J. and Christensen, P. 1956. "A factor-analytic study of verbal fluency". University of Southern California Psychology Laboratory Report No. 17. Los Angeles: University of Southern California.

Holzinger, K. and Harmon, J. 1938. "Comparison of two factor analyses". Psychometrika 3: 45-60.

Karlin, J. 1942. "The factorial isolation of the primary auditory abilities". Unpublished Ph.D. dissertation, University of Chicago.

Leutenegger, R. and Mueller, T. 1964. "Auditory factors and the acquisition of French language mastery". Modern Language Journal XLVII: 141-146.

Leutenegger, R., Mueller, T., and Wershow, I. 1965. "Auditory factors in foreign language acquisition". Modern Language Journal XLIX: 22-31.

Oller, J. Jr. 1979. Languge tests at school: a pragmatic approach. London: Longman.

Oller, J. Jr. and Hinofotis, F. 1980. "Two mutually exclusive hypotheses about second languge ability: indivisible or partially divisible competence". In Oller, J. and Perkins, K., (eds). Research in language testing. Rowley, Massachusetts: Newbury House.

Pimsleur, P., Stockwell, R., and Comrey, A. 1962. "Foreign language learning ability". Journal of Educational Psychology 53: 15-26.

Purcell, E. and Suter, R. 1981. "Predictors of pronunciation accuracy: a reexamination". Language Learning 30: 271-288.

Seashore, C., Lewis, D., and Staetveit, J. 1939. Seashore measures of musical talents. New York: The Psychological Corporation.

Suter, R. 1976. "Predictor of pronunciation accuracy in second language learning". Language Learning 26: 233-253.

Upshur, J. and Homburg, T. 1983. "Some relations among language tests at successive ability levels". In Oller, J. Jr. (ed). Issues in language testing research. Rowley, Massachusetts: Newbury House.

Wanner, E. 1974. On remembering, forgetting, and understanding sentences. The Hague: Mouton.

Wepman, J. 1973. Auditory discrimination test. Los Angeles: Western Psychological Services.

FOREIGN LANGUAGE NEEDS: THEORY AND EMPIRICAL EVIDENCE IN CZECHOSLOVAKIA

Jan Prucha (Comenius Institute of Pedagogy, Praha)

Abstract

Foreign language needs is a new interdisciplinary area of research developing on the border between applied linguistics, sociolinguistics and educational planning and evaluation. What actually is the phenomenon called FL needs? Generally speaking, FL needs can be defined as demands, wishes, expectancies etc. of the whole society, its groups and individuals reflecting the relation of their contemporary as well as future existence to communication in foreign languages. Having presented this general definition I shall proceed as follows. First, I shall illustrate how the FL needs of a society are being realized in the case of Czechoslovakia through the system of FL teaching. Then I shall describe some types of empirical studies of FL needs. The aim here is to show how FL needs can be assessed by applying various research techniques. In conclusion I shall turn to theory: I shall briefly discuss some crucial problems in the theory and methodology of research into FL needs.

1 FL needs as reflected in the educational system of Czechoslovakia

Any society has certain FL needs reflecting its geographical, political, economic, cultural, etc. conditions as well as its own (native) language and its relation to languages of other nations. It follows that it is possible, in principle, to draw up a "FL needs profile" of any society. Probably the most evident characteristic of this profile is how FL needs are met by the educational system of the society. Using this approach, I shall now briefly characterize the system of FL teaching in Czechoslovakia.

1.1 General characteristics

The Czechoslovak Socialist Republic (CSSR), with a population of 15 million, has two official languages:

- Czech (a native language of 64 % of the population), and
- Slovak (spoken by 31 % of the population).1)

Due to historical, geographical and other reasons FL teaching has a strong tradition in Czechoslovakia. At present, various FL courses are incorporated in curricula

of all types of schools (beginning from the 5th grade of the comprehensive school up to the university and post-graduate level).

All this has led to a relatively high proportion of people in Czechoslovakia knowing at least one foreign language. According to Szepe (1980) a knowledge of 4 "world foreign languages" in Czechoslovakia is much higher than in two neighbouring countries (Hungary, Poland) but somewhat lower than in Sweden, for example, or The Netherlands (see Table 1-2, App. 1). According to Szepe's data, Czechoslovakia (together with Belgium, Norway and Sweden) belongs to the group of European countries having the second highest "foreign language saturation" among inhabitants (the first place in Europe is taken by only one country - The Netherlands).2)

The knowledge of foreign languages in Czechoslovakia is primarily a result of the FL teaching in schools, the present-day system of which is as follows (cf. also information in App. 2).

1.2 Foreign languages in Czechoslovak schools

Russian is the 1st compulsory foreign language in all types of schools in Czechoslovakia beginning from the 5th grade of the basic (i.e. the 8-year comprehensive) school and extending up to university level. Other foreign languages were for the most part not taught in basic schools (with one exception - see below). However, with the contemporary reform of the educational system in Czechoslovakia a new arrangement has come into force.

Beginning from the school year 1982-83 all pupils in grades 7 and 8 of the basic comprehensive school have a right to choose a 2nd foreign language (English, French or German) as one of the elective (but compulsory) subjects. There are at present no data reflecting the interests of pupils for learning a 2nd foreign language.

Within the framework of comprehensive education, certain basic schools operate as schools with enlarged instruction in foreign languages. In 1980 there were 104 such schools with 27,376 pupils in Czechoslovakia. The curricula of the schools involve as the obligatory subject one of the following languages: English, French, German or Russian. The respective data are given in App. 2.

The 2nd foreign language is taught in all types of upper secondary schools in Czechoslovakia, i.e. in the "gymnasium" (a 4-year general secondary school) and in various types of secondary vocational schools. English, French, German or (exceptionally) Spanish are taught as compulsory subjects, and as an optional subject (in gymnasium only) Latin. This means that in every year about 38 % of the total young population learns obligatorily at

least 2 foreign languages (Russian + one of the above-mentioned foreign languages). According to some estimates the preference for the 2nd foreign language is as follows: 1. German, 2. English, 3. French (a very small proportion), 4. Spanish (taught only in some commercial secondary schools).

There is, at present, a growing tendency to emphasize the position of the 2nd foreign language in upper secondary schools through certain new arrangements; for example, by introducing a possibility (now being experimentally tested in some schools) for pupils to choose the 2nd foreign language as one of the subjects in the School-Leaving Examination (up to now Russian has been an obligatory subject in all School-Leaving Examinations).

Besides the "normal" educational system there are also public Schools of Foreign Languages (mainly evening courses) for adults who want to learn a foreign language on the basis of their private interests. In 1980 there were 35 schools of foreign languages with 62,339 learners studying English and German (cf. Table 3, App. 3). In addition, there also exist a great number of foreign language courses in cultural institutions, business, etc. as a form of out-of-school education. German and English are the most frequently taught languages there.

There is also a special form of foreign language teaching - private lessons. Some foreign language teachers provide private instruction to those who prefer more intensive and individualized foreign language teaching than that obtainable in schools and public foreign language courses. There are no official data about this form of teaching but English and German are apparently the most frequent languages taught.

At the university level Russian (obligatory) and one of the 2nd foreign languages (English, French, German and - in special fields of study - Spanish and Latin) form part of compulsory study. Of course, as regards special philological study, a number of foreign languages are taught in faculties of philosophy, pedagogical faculties and in the Institute of Sports Education. A distribution of foreign languages taught in linguistic branches (in the Czech Socialist Republic only) is given in Table 4, App. 4.

2 Empirical findings on FL needs in Czechoslovakia

In recent years some empirical studies (with a different explanatory power) on FL needs have been carried out in Czechoslovakia. They have been concerned with one of three lines:

- The actual use of foreign languages by inhabitants of Czechoslovakia
- Interest in FL learning in certain population groups
- Expert evaluation and forecasting of future FL needs

I will give a brief account of a number of studies, not with any view to producing a complete survey, but to illustrate the types of problems that have been studied and the kinds of methods applied.

2.1 Actual use of foreign languages in Czechoslovakia

There is, generally, a lack of empirical evidence concerning the actual present-day use of foreign languages by inhabitants of Czechoslovakia. The most extensive investigation so far was carried out in 1965 (Chodera and Berouska, 1968). The investigation was organized by the state administration among a representative sample of 20,428 inhabitants (aged 15-64 years) taken from all areas of Czechoslovakia. From this basic sample a group of people were identified who had actually used a foreign language in real communicative situations (N = 5,462) and investigated in detail. A special questionnaire was completed by individual respondents during a personal visit by investigators.

The main research problems were:

- What foreign language has been used by respondents and for what communicative purposes
  (a) in their professional sphere in Czechoslovakia (9 questions),
  (b) outside their professional sphere in Czechoslovakia (11 questions).

- What was the continuity and regularity of the use of the foreign language (4 categories of time were differentiated).

- What were the topics of FL communication (24 topical categories differentiated).

These problems were investigated in respect to certain characteristics of the inhabitants (sex, age, professional status, place of residence). I shall mention only some of the most relevant findings here:

The distribution of the foreign languages used for various communicative purposes in all areas of Czechoslovakia was:

1. German (43,6 % of all communicative acts),
2. Russian (29,1 %),
3. English (14,5 %),
4. French (10,9 %),
5. Spanish (1,9 % - other foreign languages were not considered).3)

What is worth mentioning is that, generally,

(a) there were no significant differences in the rank order of use of particular foreign languages with respect to regional areas, i.e. Czechoslovakia is a homogenous country in this sense,

(b) there were significant differences between the particular foreign languages concerning the use for various communicative purposes, topics of communication and regularity of communication. For example, the most frequent communicative activities were: in German - ordinary social conversation; in Russian - oral communication for specific working purposes; in English - reading and translating texts for specific working purposes. The particular age groups and categories of occupation also, of course, showed different patterns of foreign language use.4)

This investigation was concerned with the population of Czechoslovakia as a whole. There is, however, a lack of empirical studies dealing with use of foreign languages in specific occupational groups. I have recently analyzed the use of foreign languages in several branches of scientific research in Czechoslovakia (Prucha, 1975). In this study, two technical disciplines (electronics, metallurgy) were compared with two humanistic disciplines (linguistics and pedagogy). The aim of the study was to ascertain what information sources (with respect to foreign languages) are utilized in the disciplines concerned (the study was based on citation analysis). The main results are given in Table 5, App. 3.

It can be seen that researchers in the given four sciences in Czechoslovakia use mainly English and Russian when drawing from information sources in the respective fields, and to a certain degree German and (less frequently) French. In particular, for pedagogical research the following distribution was found: 1. Russian (48.4 % of all information sources), 2. English (22.8 %), 3. German (15.3 %), 4. Latin (5.4 %), 5. Polish (4.8 %), 6. other foreign languages (3.3 %).

These are some data concerning the use of foreign languages in Czechoslovakia. How can one interpret the data? Do they reflect actual FL needs or not? This is not a simple question; I shall discuss it in more detail in Part 3.

2.2 Interests and motivation for foreign languages

For research purposes it is necessary to compare the actual use of foreign languages in a society with individuals' needs for foreign languages revealed in their interests, attitudes and motivation for learning foreign languages. Concerning the latter, an important study was carried out by Pychova (1975) analyzing the choice-preferences for modern foreign languages in a sample (N = 826) of gymnasium students in the industrial district of Ostrava. In the sample the highest preference was for English (65.9 % of students), then German (27.8 %), and the lowest preference for French (6.3 %) - in all cases as the 2nd foreign language. The detailed analysis concentrated on reasons the students gave for motivating their foreign language learning. The most frequent reasons were: a need for personal communication with foreigners, especially when travelling abroad; and a need for reading and understanding a special (scientific) literature. However, the choice of English was often motivated by the popularity of the language created through pop-music, films, etc. If the parents influence their children's choice of a 2nd foreign language they prefer German, while the pupils without parental influence prefer English.

In my opinion, this is an indication of an interesting phenomenon, namely a developmental shift in FL needs in the Czechoslovak population. The young generation (upper secondary and university students) prefer English to German while adult learners reveal the contrary preference, i.e. German to English (which is in correspondence with the general use of the two languages in Czechoslovakia). Does it mean that adults are more adequately aware of actual FL needs? This developmental change has not been studied as yet.

This shift in choice of a foreign language was also evidenced in another analysis by Pychova (1975), namely in a sample of adult learners at the School of Foreign Languages in the same district of Ostrava. Among 3,163 learners (studying a foreign language voluntarily in evening courses) some 52 % chose German while only 38.6 % English. At the same time the motivation types are significantly different: the main reason for learning German is the need for personal communication with foreigners while in the case of English the main reason is to study special literature for working purposes.

Another investigation was carried out by Hašpl (1978) with a sample of adult learners at the School of Foreign Languages in Prague.5) The aim of this study was to find the motivations (reasons) for foreign language learning in a sample of 224 English learners. The average age of the sample was 34 and a large proportion of the learners had finished university (34 %) or upper secondary school (52 %).

The following main reasons (i.e. individual needs) for learning English were ascertained:

- to use English when travelling abroad (51 %),
- to study special literature (48,6 %),
- not to lose an ability to learn (45.7 %),
- to use English in employment (42.3 %).

Hašpl (1978) interpreted his findings as indicating that within individuals' FL needs a significant role is played not only by "material needs" (i.e. concerning the use of foreign language in employment where, in many cases, a certificated knowledge of a foreign language leads to a higher reward) but also by "needs for self-development" (i.e. needs for cultural development of one's own personality as revealed, for example, in the desire to learn about the culture and life of foreign nations, etc.).

However, in another study (Vackova, 1967) carried out on a sample (N = 1,600) of adults learning Russian at the same School of Foreign Languages in Prague quite another hierarchy of motives was found: 45 % of learners were motivated through "material needs" (i.e. learning Russian for purposes of professional work) while only 23.5 % of learners gave reasons of "cultural self-development" (i.e. interests in Soviet culture, travelling in the Soviet Union, etc.).

Does it mean that individuals' FL needs are differentiated in respect of particular languages? This phenomenon should be studied more thoroughly before an answer can be given. On the other hand, there is no research in Czechoslovakia concerning the fact that a knowledge of a foreign language tends to increase the social prestige of an individual.

Some studies in Czechoslovakia have used expert opinions to obtain data on the future importance of particular foreign languages for educational purposes. For example, Zlamalova (1980) presents a prediction that with respect to the long-term needs of Czechoslovak society the following distribution of foreign languages would be optimal at upper secondary level: English (40 % of students should learn the language), German (40 %), French (15 %), Spanish (15 %). (Russian was not ranked as it is assumed that it will remain the 1st foreign language in

schools also in the future.) This estimation is, however, far from the contemporary state of affairs (cf. data in Part 1) where the proportion of French and Spanish learners is very low and shows no sign of increasing.

3 Theoretical problems of research into FL needs

The notion "FL needs" remains insufficiently explained in certain respects. This is perhaps a consequence of the fact that in the research into FL needs (not only in Czechoslovakia but in other countries, too - cf. a survey by Oud-de Glas, 1982) the empirical, descriptive approaches have dominated while a theoretical analysis of the phenomenon has not developed.

In a monograph (Prucha, 1983) I have tried to contribute to a theoretical explanation of the "needs of society" related to language teaching. Basically, I have applied the general conception of needs maintained by Marxist philosophy in Eastern Europe. A general concept of needs has been defined in the following manner: a need is the state of an organism, human being, social group or whole society, which reflects the dependency of their existence on objective conditions and, at the same time, a need is a source of activities aimed at changing the conditions (Filosofskaja enciklopedija, 1967).

It follows that it is necessary to distinguish two basic levels of needs:

a) individual needs,
b) needs of society.

Accordingly, I am treating individual FL needs as those needs which are reflected in motivations, interests, aspirations, expectations, attitudes of individuals with respect to foreign language learning. A typical example is that an individual chooses which foreign language he wants to acquire in order to fulfil his own need.

The FL needs of society are those needs which are reflected in decisions of state authorities (of educational policy, etc.) concerning foreign languages taught in schools, their order of importance and the degree of knowledge prescribed, and which are supposed to be relevant for the whole society. A typical example is that the curricula of certain schools involve one or more foreign languages whose teaching is compulsory for certain groups of the young population.

Individual FL needs and the FL needs of society may coincide or (as more often occurs) may be more or less in contradiction. A contradiction may result in the following ways:

- The FL needs of society as codified by documents of educational policy accurately reflect real needs but the FL needs of individuals are oriented in another way. In the Czechoslovak situation this is the case with e.g. French: for upper secondary schools to offer courses in French is desirable for the society, but the choice of French by students has declined to a very low proportion.

- The FL needs of society as codified by educational documents do not reflect real needs and also the FL needs of individuals are oriented in another way. In the Czechoslovak situation this is the case with e.g. Spanish and Latin, which are not involved as compulsory subjects in all types of upper secondary schools and which also arouse little interest among students.

A number of research problems arise concerning individual FL needs and, above all, the FL needs of society. I shall mention some of them.

3.1 How can the FL needs of society be objectively identified?

This means: From what sources of the social reality and by means of what methods can we identify FL needs? The problem is concerned not only with solutions as regards the range of foreign languages which ought to be taught in schools but also with some other decisions having impact on national economy (training a desirable number of foreign language teachers, interpreters, linguists, etc.).

In my previous work (Prucha, 1983) I delimited several sources for identifying the needs of society with respect to language teaching. The main principle applied is a dynamic approach, i.e., it is necessary to study what would be the future life activities of learners currently acquiring a foreign language.6) In other words: what use the learners' knowledge of a foreign language will be in the future. It is very difficult to develop such predictions. Recently, however, certain methods have been applied which make it possible to forecast the future activities of present-day pupils and, accordingly, to modify aims and contents of foreign language courses so that they can meet the future FL needs of pupils. This matter has been dealt with in the theory of "didactic forecasting" (Geršunskij and Prucha, 1979).

3.2 What is the real substance of FL needs?

The problem can be answered adequately on the operational level only, i.e. through examining human activities in which FL needs reveal themselves. A typical activity, for example, is the selection of a foreign language to learn. The selection is probably motivated by, among other things, the (real or estimated) importance of a foreign language. Again, however, one can ask how we can identify

the importance of foreign languages. In my previous research I have applied four general criteria of importance of foreign languages (with respect to Czechoslovak society):

- the population extent of the foreign language concerned (i.e. the number of people using the language as a mother language),
- the geographical extent of the foreign language (i.e. the number of countries and the size of their territory where the language is in use),
- the use of the foreign language in international communication (e.g. the use of the foreign language as a means of communication in international trade, research, diplomacy and tourism),
- the difficulty of the foreign language for Czechs and Slovaks (i.e. how difficult it is to learn as measured by the degree of its genetic relationship with the Czech or Slovak language).

The criteria are, of course, not exhaustive and have a different explanatory power with respect to the individuals's choice of which foreign language to learn. For example, it remains unclear to what degree the difficulty of a foreign language influences the selectional process: in some cases an individual takes into consideration the supposed difficulty of a foreign language and, accordingly, modifies his FL need, but in other cases not.

Certainly, there is a need to define criteria by which to assess the importance of foreign languages (for individuals as well as for the whole society) and its impact on FL needs. This is, of course, only one among a number of tasks stemming from research into FL needs (cf. a schematic survey in App. 5).

Notes

1 Czech and Slovak are closely cognate Slavonic languages. This relationship causes a special kind of bilingualism among Czechs and Slovaks: chiefly passive bilingualism, i.e. without active command of the cognate language but with almost full mutual understanding on both sides (for details see Salzmann, 1980). The largest language minority in Czechoslovakia is Hungarian (almost 4 % of the population).

2 There exist, however, no exact data about the degree of real knowledge and efficient use of foreign languages in the countries named. The data given by Szepe (1980) perhaps somewhat overestimates the actual knowledge of foreign languages.

3 The data reflect the situation in the year 1965. For the present-day situation I would estimate the same order of foreign languages but with some changes in proportion of use: an increase in the case of English and a decrease in the case of German and French.

4 It is no surprise that German holds the first place for communicative use among other foreign languages in Czechoslovakia. Let us compare, for example, the data concerning the number of foreign visitors coming every year to Czechoslovakia. According to the Statistical Year-Book of the Czechoslovak Socialist Republic (Statisticka ročenka Československe socialisticke republiky, 1981) in 1980, of the total 17,362,752 foreign visitors 43,5 % came from German speaking countries (GDR, FRG, Austria) while only 1,4 % came from the USSR. If one considers the fact that visitors from other countries (Switzerland, Hungary, etc.) traditionally use German as a means of communication in Czechoslovakia, it becomes still more apparent that German is the foreign language most frequently met by Czechoslovak inhabitants.

5 The School of languages (Jazykova škola) in Prague is the largest institution for teaching foreign languages in Czechoslovakia with more than 17,000 learners studying in various courses of some 30 foreign languages.

6 On this point, cf. a discussion on future FL needs of pupils in A.L. Leino (1978), pp. 21-24.

References

Chodera, R. - Berouška, J.: Výberove setreni o používani cizich jazyku (Sample Study into Foreign Languages Use). In: Teoreticke otazky ruske konverzace. Praha, SPN 1968, pp. 97-127.

Filosofskaja enciklopedija (Encyclopedia of Philosophy), ed. by F. K. Konstantinov. Moskva, Nauka 1967.

Geršunskij, B. S. - Prucha, J.: Didaktičeskaja prognostika (Didactic Forecasting). Kiev, Višca škola 1979.

Hašpl, F.: Motivace učastiniku studia cizich jazyku (Motivation of Foreign Language Learners). Praha, Filozoficka fakulta UK 1978 (unpublished thesis).

Leino, A. L.: An Introduction to the Content of Foreign Language Instruction. Research Bulletin No. 50, Institute of Education, University of Helsinki, Helsinki 1978.

Oud-de Glas, M.: Foreign Language Needs: A Survey of Research. Nijmegen, Institute for Applied Sociology 1982.

Prucha, J.: Využivani informaci v československem pedagogickem vyzkumu (The Use of Information Sources in Czechoslovak Pedagogical Research). Pedagogika, 25, 1975, No. 4, pp. 461-477.

Prucha, J.: Potřeby socialisticke společnosti a vzdělani (Needs of the Socialist Society and Education). Praha, Academia 1983.

Pychova, I.: Studie o motivaci posluchaču jazykovych škol a gymnasii k učeni cizim jazykum (Study on Motivation for Learning Foreign Languages). Ostrava 1975 (unpublished research report).

Salzmann, Z.: Language Standardization in a Bilingual State: The Case of Czech and Slovak. Language Problems and Language Planning, 4, 1980, No. 1, pp. 38-54.

Statisticka ročenka Československe socialisticke republiky 1981 (Statistical Year-Book of the Czechoslovak Socialist Republic 1981). Praha, SNTL 1981.

Statistika školstvi 1980-1981 České socialisticke republiky (Statistics of the School System 1980-1981 of the Czech Socialist Republic). Praha, Ustav skolskych informaci 1981.

Szepe, G.: Less Taught Languages in Europe (Their Place in Education and Their Role). Paris, UNESCO Division of Structures, Content, Methods and Techniques of Education, Jan. 1980.

Vackova, J.: Proč se učí ruštine (Why They Learn Russian). In: Prace, Sept. 5, 1967.

Zlamalova, M.: Ktery z volitelnych cizich jazyku na našich školach je nejdulezitejsi? (Which of the Optional Foreign Languages in Our Schools is the Most Important?). Cizi jazyky ve škole, 24, 1980-81, c. 2, pp. 82-83.

Appendix (1)

Table 1

Proportion of population with a knowledge of "world" FL

Country	English	French	German	Russian
Czechoslovakia	10.4 %	2.0 %	23.6 %	27.7 %
Hungary	9.7 %	1.9 %	14.5 %	19.4 %
Poland	7.5 %	3.0 %	9.1 %	6.0 %
Romania	4.9 %	19.8 %	7.4 %	4.9 %
Yugoslavia	6.3 %	1.0 %	6.3 %	8.3 %
Netherlands	83.9 %	7.6 %	22.9 %	-
Sweden	50.0 %	6.2 %	25.0 %	-

Table 2

Proportion-classes (according to the number of people who speak at least one of the four "world" FL)

maximum ↑

(A) Netherlands
(B) Malta
(C) Belgium, Czechoslovakia, Norway, Sweden
(D) Bulgaria, Finland, Hungary, Romania
(E) Austria, Germany (DDR, FRG), Poland
(F) Spain, Yugoslavia, USSR
(G) France, Greece
(H) Turkey
(I) Portugal, UK (England rather)

Source: Proportions calculated on the basis of data published by Szepe (1980).

Appendix (2)

Information about FL teaching in the compulsory school system in Czechoslovakia

Compulsory education for all from age 6 (10 years of education obligatory):
grade 1 - 8 (basic comprehensive school), afterwards:
grade 9 - 12 (gymnasium or other type of upper sec. school)
or grade 9 - 10 (11) (schools for vocational education)

Foreign languages taught:

Basic school
A-language: Russian (obligatory for all from grade 5)
B-language: English or German or French (optional in grade 7-8)

Upper sec. school (gymnasium)
A-language: Russian (obligatory for all in all grades)
B-language: English or German or French (obligatory for all in all grades, language elective)
C-language: Latin (optional in all grades)
D-language: one B-language not studied as a B-language (e.g. if English is chosen as B then either German or French can be studied as D)

Distribution of teaching hours per week (length of hour: 45 min):

		A-language	B-language	C-language	D-language
basic school	grade 5	4	-	-	-
	6	3	-	-	-
	7	3	2	-	-
	8	3	2	-	-
gymnasium	grade 9	3	3	2	2
	10	3	3	2	2
	11	3	3	2	2
	12	3	3	2	2
total hours (35 weeks/year)		875	560	280	280

Appendix (3)

Table 3

Foreign languages taught in Schools of Foreign Languages in Czechoslovakia

language	number of learners	proportion in %
German	30,052	48.0
English	23,469	37.5
French	4,071	6.6
Russian	1,731	2.7
Spanish	1,007	1.6
other languages1)	2,309	3.6
total	62,575	100.0

1) Among other languages the most taught ones are Italian, Serbo-Croatian, Swedish and Japanese.

Source: Statistika školstvi 1980-1981 České socialisticke republiky (1981).

Table 5

Actual use of foreign language information sources in some fields of research in Czechoslovakia

metallurgy	electronics	linguistics	pedagogy
1. Russian	English	English	Russian
2. English	Russian	Russian	English
3. German	German	German	German
4. French	French	French	French

Source: Prucha (1975).

Appendix (4)

Table 4

Foreign languages taught at universities (philological line) in the Czech republic in 1980

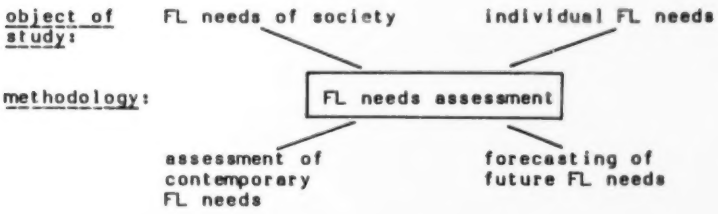
language	number of students
Russian	1,978
German	422
English	300
French	61
"oriental" languages1)	80
Spanish	20
Slavonic languages2)	20
Latin	22
other foreign languages3)	no data

1) Arabic, Chinese, Japanese, Vietnamese and some other languages of Africa and Asia.
2) All Slavonic languages except Czech, Slovak and Russian.
3) No data available about the number of students. Foreign languages taught are: Danish, Dutch, Finnish, Greek, Hungarian, Italian, Norwegian, Portuguese, Swedish.

Source: Statistika školstvi 1980-1981 České socialisticke republiky (1981).

Appendix (5)

RESEARCH INTO FL NEEDS (EDUCATIONAL ASPECT)



research approaches:

(1) data collection (questionnaire) on FL needs among FL users and other informants (e.g. employers)
- about real use of FL for professional and other specific purposes,
- about importance of particular FL (e.g., in industry, commerce, tourism, etc.),
- about required level of FL skills, etc.

(2) analyzing FL curricula at certain types of schools, e.g. what FL courses are offered in schools, the number of lessons given, the number of pupils participating in FL courses, the content of FL curricula, etc.

(3) analyzing reasons, attitudes, interests, prejudices, etc. of students and ex-students which determine their choice for FL, e.g.
- motives for and attitudes towards learning of FL,
- individual's assessment of importance of FL,
- opinions of ex-students as to whether FL teaching they had meets their actual demands, etc.

fields of application: mainly in educational sphere, e.g., in:
- developing recommendations for national FL educational policy,
- evaluating projects of FL curricula,
- evaluating efficiency of FL teaching programmes in operation,
- influencing FL needs of individuals through educational propaganda,
- influencing cooperative language policy on international level, etc.

INVESTIGATING DIMENSIONALITY OF LANGUAGE TESTS - A NEW SOLUTION TO AN OLD PROBLEM

Ulrich Raatz (University of Duisburg)

1. Introduction

The content of this paper is factorial validity of language tests. Two of the most interesting questions in this area are the questions about the dimensionality of language and about the dimensionality of language tests.

These two questions and the methods of language teaching are closely connected. A given structure demands specific measurement techniques and teaching procedures. Thus a structuralist model of language with the components listening, speaking, reading and writing will produce listening comprehension tests, oral tests, reading comprehension tests and writing tests. An integrated approach demands global (holistic) language tests such as the interview, the essay or the cloze test.

At the present time these two models of language are in competition with each other. Each is based on a theory of language, each is confirmed by empirical data, each has led to the development of specific testing procedures.

Which model is correct? Or, perhaps, to put it more accurately, which model is a better description of reality?

Perhaps it is possible to solve the problem on the basis of theoretical considerations alone. Another approach, however, would be to use empirical methods.

Such an empirical examination of the proposed hypotheses about the structure of language tests, and consequently also about the structure of language itself, is generally conducted by administering a number of different language tests to a sample of subjects and performing factor analysis on the test results. The factors discovered enable conclusions to be drawn in favour of one of the above-mentioned models.

This type of research design can be illustrated by a number of studies. The results of these studies cannot be said to be unanimous. They differ in the existence of a general factor, in the number of common factors, and in the interpretation of these. In short, they do not help us to decide which of the hypotheses about the structure of language is the "correct" one. How can this lack of unanimity be explained?

Detailed examination of the studies reported above shows considerable differences in the methods and the evaluation techniques

used. Some of the samples are homogenous, some are heterogenous. Each study uses different tests. But the main problem is the use of factor analysis. This statistical method is problematical for solving the problem of dimensionality. Why?

2. A brief overview of factor analysis

Factor analysis is a method belonging to the area of descriptive statistics. It permits the researcher to reduce large quantities of information (for instance the results derived from a large number of tests) to certain common underlying aspects or dimensions; the so-called factors.

A factor analytic study using the methods originally developed by Thurstone (1947) normally proceeds in the following steps:

1. A number of tests are selected which are to be described more parsimoniously, namely by reduction to a small number of underlying factors.

2. A sample of examinees is selected.

3. The tests are administered to this sample.

4. All intercorrelations between the tests are calculated.

5. The individual factors are "extracted", in other words the relevant loadings are determined for one factor after another. For this problem there are a number of different methods available.

6. The number of factors finally extracted is determined by a so-called cut-off criterion (e.g. Eigenvalue-criterion).

7. Rotation is performed. This is done because factor analysis has an infinite number of equally satisfactory solutions which can all be converted into each other by rotation. Every single solution is correct.

   During rotation the number of factors generally remains constant; what changes is the amount of variance explained by the individual factors. So a g-factor can turn into a group factor.

   The criteria used to determine which of the infinite number of possible rotations is the "correct" one can be either psychological (intuitive) or analytic. A researcher using psychological criteria selects the solution which best supports his original hypothesis. Analytic rotations are determined by mathematical methods. The best-known of these is the varimax rotation to "simple structure".

8. Finally the factors are named. Generally the names chosen are determined by the tests which have the highest loadings on the individual factors.

3. Some problems with factor analysis

Factor analysis has a number of weak points so that it is possible to manipulate the results. The most important of these are:

1. the selection of the sample of examinees
2. the selection of the tests
3. the method of extraction chosen
4. the method of rotation chosen

ad 1: Kalveram (1965,1970) showed theoretically that the results obtained by a factor analysis could be dependent on the sample. If a sample is extremely heterogeneous one will tend to find only few factors, in extreme cases only a g-factor. If, on the other hand, the samples are homogeneous more factors will emerge.

ad 2: The reason for this is obvious. If, for instance, no oral test is included in the test battery, a factor analysis of the battery cannot yield an oral factor.

Only if a number of similar oral tests are added to the battery, and thus determine a large part of the variance in the matrix, is it probable that an independent oral factor will emerge. Obviously, then, the detection of certain factors is the result of the selection of tests entered into the factor analysis.

ad 3: The use of different methods of extraction of the factors can produce different factor patternings. This is true for the most commonly used common factor method, for the principal component method,and the method invented by Spearman.

ad 4: The rotation performed after extraction has taken place is a matter of personal choice. There are an infinite number of correct solutions. Thus, a "psychological" rotation can be performed, aimed at producing an optimal approximation to a particular theoretically-determined structure. Even if analytic rotation is used, there are a number of possibilities available, some of which tend more to a g-factor, some more to group factors.

Perhaps these weak points account for the different results of factor analytic studies of dimensionality of language tests. In fact, no researcher will manipulate the results, but everyone works in a different manner. You see, that factor analysis is not the best method for solving the problem of language dimensions.

4. The CLA-model

I propose a different solution. This is provided by the CLA-model proposed by Moosbrugger and Müller first in 1980 (1982). CLA stands for Classical Latent Additive Test Model, and, as the name implies, this is a latent-trait model, comparable to the Rasch-Model.

In its original form, this CLA-model is applicable to items which are least ordinally scaled. I have examined the use of this model for tests, in particular for investigating whether a number of tests are undimensional or not.

The CLA-model has the same characteristics as the Rasch-model:

- the true score is split into an ability parameter and a difficulty parameter;
- it has what is called specific objectivity; that means that the item difficulties and subject abilities can be estimated independently of each other;
- if the model fits, then the total score is undimensional and interval-scaled;
- model fit can be tested empirically.

I shall now present the basic characteristics of the model and illustrate its use with three examples.

The CLA-model has the following basic assumptions:

A test consists of c independent, equivalent parts (items, superitems, subtests). X_{pi} is the raw score of subject p (p = 1...N) in test part i (i = 1...c).

Just as in classical test theory we have the assumption that

$$X_{pi} = \tau_{pi} + \varepsilon_{pi} \qquad (1)$$

where τ is the true score and ε is the error component. The second assumption is

$$E(\varepsilon_{pi}) = 0 \qquad (2)$$

It follows, therefore, that

$$E(X_{pi}) = \tau_{pi} \qquad (3)$$

i.e the expected value for X_{pi} is the true score τ_{pi}.

In the CLA-model we split the true score into two components:

$$\tau_{pi} = \xi_p - \beta_i \qquad (4)$$

ξ_p is the ability of subject p; β_i is the difficulty level of item i.

Just as in the Rasch model we assume local stochastic independence, both on the superitem level (independence of content) and on the subject level (no copying).

From these assumptions we can now derive the "specific independence" of the model.

If we take two persons p and q and an arbitrary item i, then it follows for the difference of the two expected X-values:

$$E(X_{pi}) - E(X_{qi}) = \xi_p - \xi_q \qquad (5)$$

This difference is independent of the item.

If we take one arbitrary person p and two items i and j, then it follows for the difference of the two expected X-values:

$$E(X_{pi}) - E(X_{pj}) = \beta_j - \beta_i \qquad (6)$$

This difference is independent of the person.

The item characteristics in this model have the following form:

Figure 1

Item characteristics in the CLA-model

The model parameters ξ and β can be estimated as follows:

$$\xi_p' = \frac{1}{c} \sum_i (X_{pi} + \beta_i) \qquad (7)$$

or, if we normalise the β-values so that their sum equals 0, we can simplify the equation to:

$$\xi_p' = \frac{1}{c} \sum_i X_{pi} \qquad (8)$$

For the items we have:

$$\beta_i' = \frac{1}{N} \sum_p (\xi_p - X_{pi}) \qquad (9)$$

In this formula ξ_p has to be estimated by the values derived from (8).

In the case of a test battery with c subtests with n items each these formulae can be simplified. Let X_{pi} be the raw score of the subject p in subtest i, X_{ptot} the total score, $\bar{X}_i$ the mean score in subtest i and X_{tot} the overall mean. Then we find from (8) that:

$$\xi_p' = \frac{1}{c} X_{ptot} \qquad (10)$$

and from (9)

$$\beta_i' = \bar{X}_{tot} - \bar{X}_i \qquad (11)$$

There are a number of ways in which one can check for model fit with the CLA model:

1. The subject group is split into two parts using some arbitrary criterion. For each group the item parameters (β's) are calculated separately. The pairs of betas (one for each group) should show identical values which have only chance differences. For each pair the deviations are tested with the t-statistic; for the items overall with the chi-square statistic.

2. The item characteristics for all items should be parallel. This means, of course, that each superitem must have the same variance and item discrimination index. This condition is satisfied if TUKEY's test for non-additivity is non significant.

If these model checks turn out to be negative, the items or subtests which do not fit the model can be easily identified and

eliminated. If the model fits, the sum scores of the items, the superitems or the subtests are assumed to be undimensionally scaled on an interval scale.

5. 1st example: C-Test

In Duisburg a project is currently in progress which is supported by the "Arbeiterwohlfahrt", a social service agency. The aim of this project is to enable gifted Turkish girls to obtain a higher educational qualification. Around 30 girls attend courses for two years, and at the end of this time they take the "Mittlere Reife" examination. The number of applicants for this programme is much larger than the number of places available, and therefore selection procedures are necessary. Former school grades are the one selection criterion; the other is proficiency in the second language, German. This proficiency has, up to now, been assessed on the basis of a home-made test. The people in charge of this project asked us whether we could develop a German C-Test which would either provide additional information or even replace the present test procedure.

We constructed a C-Test with 5 parts. Each part had 20 deletions.

26 Turkish girls took the tests at the first test session in March 1984.

If we use the CLA-model on the C-Test data for test evaluation we begin by deriving the beta values (eq. 9) given in Table 1 as item difficulty indices for the C-Test parts.

Table 1

CLA-difficulties of the 5 superitems and the tests for goodness of fit

	β	β_{sup}	β_{inf}	t-test
C 1	-5,6	-5,0	-6,2	1,51 n.s.
C 2	-0,6	-0,7	-0,4	0,55 n.s.
C 3	2,9	2,7	3,2	0,98 n.s.
C 4	1,1	0,8	1,4	0,84 n.s.
C 5	2,1	2,2	1,9	0,45 n.s.

$\chi^2 = 4,45$ $(50\% > p > 25\%)$
Tukey: $F = 10,3$ $(p < 0,2\%)$

For the first test of model fit we separate the group on the basis of their C-Test scores into a higher and a lower half. Then the item parameters are determined for each group separately. These must be identical if the model fits.

The results are given in Table 1, together with the results of the t-tests. As one can see there is no pair for which the differences in the beta values are statistically significant, and the chi-square test over all values is not significant.

But the second model fit is unsatisfactory since the TUKEY test was statistically significant ($p < 0,2\%$). It is obviously superitem 1 which is causing problems, as inspection of the t-tests shows. We therefore decide that item 1 ought to be eliminated.

Table 2

CLA-difficulties of the 4 superitems and the tests for goodness of fit

	β	β_{sup}	β_{inf}	t
C 2	-2,0	-1,9	-2,0	0,20 n.s.
C 3	1,5	1,3	1,8	0,92 n.s.
C 4	-0,3	-0,3	-0,2	0,14 n.s.
C 5	0,7	0,9	0,4	0,84 n.s.

$\chi^2 = 1,61$ ($p > 50\%$)
Tukey: $F = 0,51$ ($p \approx 70\%$)

Table 2 shows the newly calculated item difficulties β, and the results of the two checks for model fit. Once again there are no significant individual differences, and the overall chi-square test is not significant. Now the TUKEY test is not statistically significant either. We can assume model fit. This means:

- the C-Test consisting of parts 2 to 5 is assumed to measure undimensionally;
- the total scores on the C-Test are assumed to be scaled on an interval level.

6. 2nd example: Interview

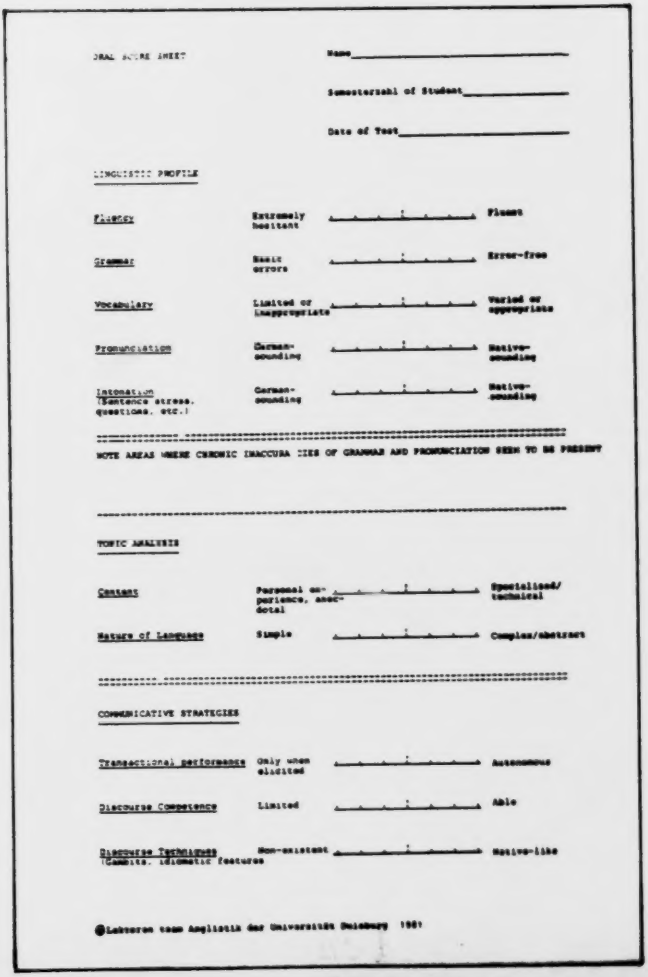
In Duisburg, the oral performance of all students of Anglistik is measured at the end of the 3rd or 4th semester. This test is a mixture of prepared rôleplay and interview in a group of 3 students with random contest and takes about 20 minutes (see Smith and

Klein-Braley, 1981, and Smith, 1982).

The randomly selected raters rate the characteristics of the oral competence of each student independently (on 6-point rating scales) using a score sheet with 10 categories.

Figure 2

Oral score sheet



ORAL SCORE SHEET

Name_____

Semesterzahl of Student_____

Date of Test_____

LINGUISTIC PROFILE

Fluency	Extremely hesitant	▲ . . . . : . . . . ▲	Fluent
Grammar	Basic errors	▲ . . . . : . . . . ▲	Error-free
Vocabulary	Limited or inappropriate	▲ . . . . : . . . . ▲	Varied or appropriate
Pronunciation	German-sounding	▲ . . . . : . . . . ▲	Native-sounding
Intonation (Sentence stress, questions, etc.)	German-sounding	▲ . . . . : . . . . ▲	Native-sounding

NOTE AREAS WHERE CHRONIC INACCURACIES OF GRAMMAR AND PRONUNCIATION SEEM TO BE PRESENT

TOPIC ANALYSIS

Content	Personal experience, anecdotal	▲ . . . . : . . . . ▲	Specialised/ technical
Nature of Language	Simple	▲ . . . . : . . . . ▲	Complex/abstract

COMMUNICATIVE STRATEGIES

Transactional performance	Only when elicited	▲ . . . . : . . . . ▲	Autonomous
Discourse Competence	Limited	▲ . . . . : . . . . ▲	Able
Discourse Techniques (Cohesion, idiomatic features	Non-existent	▲ . . . . : . . . . ▲	Native-like

©Lehrerteam Anglistik der Universität Duisburg 1981

BEST COPY AVAILABLE

In this Figure you see the 10 categories with the rating scales and the descriptions of the extreme points of these scales.

Our question was: how many dimensions does this language profile have. We performed the CLA-analysis for a group of 77 students taking the test in autumn 1982. For each student, two ratings had been given by the raters, so we calculated the mean for every scale.

The precondition was fulfilled: on each rating-scale the raw score ranged from 1 to 6 points.

The results of the CLA-analysis of the 10 scales are presented in Table 3.

Table 3

Results of CLA-analysis of oral test

	β_{sup}	β_{inf}	t-test
Fluency	-.61	-.58	.12 ns
Grammar	.03	-.19	.84 ns
Vocabulary	-.49	-.19	1.37 ns
Pronunciation	.22	-.38	3.06 hs
Intonation	.51	-.08	3.02 hs
Content	1.64	1.67	.06 ns
Nature of language	.69	.62	.30 ns
Transactional perform.	-1.19	-.74	1.63 ns
Discourse competence	-1.00	.69	1.46 ns
Discourse techniques	.20	.56	1.67 ns

$\chi^2 = 13.5$ ($P < 10\%$)
Tukey: $F = 0{,}62$ ($P \simeq 43\%$)

The first column shows the names of the scales, the second and third the estimations of the difficulty indices, beta, in the upper and lower scoring group, and the 4th the result of the t-tests. Only two t-tests, for the scales "pronunciation" and "intonation", were significant at the 0,1% level. The overall χ^2-Test was nearly significant (10 %), the TUKEY-Test for additivity was non-significant (43%). These results showed that the scale of the 10 categories was not unidimensional.

Therefore we deleted the two categories "pronunciation" and "intonation" and calcualted a second CLA analysis for the remaining 8 categories (Table 4).

Table 4

Results of CLA-analysis of oral test
with reduced categories

	B_{sup}	B_{inf}	t-test
Fluency	-.51	-.64	.47 ns
Grammar	.12	.25	1.34 ns
Vocabulary	-.39	-.25	.63 ns
Content	1.73	1.61	.21 ns
Nature of language	.78	.56	1.00 ns
Transactional perform.	-1.10	-.80	1.11 ns
Discourse competence	-.91	-.75	.76 ns
Discourse techniques	.29	.50	.98 ns

$\chi^2 = 6.5$ ($P > 50\%$)
Tukey: $F = .03$ ($P \approx 86\%$)

The t-tests, the overall χ^2-test and the test of additivity are not significant. This means that the scale consisting of the remaining 8 categories is assumed to be unidimensional. All these categories measure the same construct on an interval scale. The two categories "pronunciation" and "intonation" constitute perhaps one or two additional dimensions.

7. 3rd example: Delta-Test

In October 1984, we gave a modified form of the Delta-Test (Duisburg English Language Test for Advanced Students, Klein-Braley and Lück 1980) as a placement test to 65 beginners.

This test consisted of the following parts, each with 25 items.

Table 5

Modified Delta-Test
Duisburg, Oct 1984

Voc 1	slot-filling	active vocabulary
Voc 3a	MC 5 distractors	passive vocabulary
Voc 3b	MC 5 distractors	passive vocabulary
Gra 1	MC 5 distractors	variety of problem areas
Gra 2	slot-filling	verb tenses and aspects
Gra 3	slot-filling	prepositions and verb complem.
C ?		
C 3	5 texts	
C 4	slot-filling	integrative
C 5		
C 6		
Dict 1		
Dict 2	182 word text	
Dict 3	7 parts with	
Dict 4	25 words each	difficult formal prose
Dict 5	(beginning with the	integrative
Dict 6	8th word)	
Dict 7		
Trans 1	25 short sentences to be translated from G to E MC 4 distractors	formal prose integrative
Trans 2		
Trans 3	2 texts to be translated from G to E each with 25 parts right/wrong	formal prose integrative

The tests for grammar and vocabulary were discrete-point item tests with MC-or openended items. The holistic tests were a C-Test with 5 parts, a 183-word dictation and 3 translations, one of these in MC-format.

Our question was again: is this battery of 21 language tests unidimensional or not?

In the following Table 6 you see the results of some of the calculations we made.

Table 6

Results of CLA-analysis of the modified Delta-test

combination	number of tests	number of sign. t-tests	result of χ^2-test	result of TUKEY-test
all tests	21	12	χ^2 = 70.9 P < 1%	F = 118.4 P = 0%
voc and gram	6	0	χ^2 = 1.8 P > 50%	F = 0 P = 84%
C-tests	5	0	χ^2 = 3.0 P > 50%	F = 0.1 P = 72%
dictation	7	5	χ^2 = 20.5 P < 1%	F = 150.6 P = 0%
translation	3	2	χ^2 = 6.8 P < 5%	F = 3.7 P = 6%
optimal comb	11	0	χ^2 = 7.4 P > 50%	F = 0 P = 93%

the optimal combination consists of the following tests:
voc (3) gra (3) C-Test (3) trans (2)

In the first line you see the results of our first global analysis of all our 21 subtests. 12 t-tests are significant, and the χ^2-Test and the TUKEY-Test are also significant. This means that the 21 subtests are <u>not</u> unidimensional, in other words, that there must be more than one language factor in this test battery.

Secondly, we analysed the six discrete-point tests measuring vocabulary and grammar with the CLA-model. You see that these tests are measuring the same thing.

The 5 C-Tests also constitute only one dimension.

But, as you see, the seven parts of the dictation are heterogenous; this test is measuring at least two different dimensions of language. Also the 3 translation tests are not unidimensional.

After this, we looked for an optimal combination of subtests. We made a number of attempts and at last we found an optimal combination.

If one adds the three discrete-point tests of grammar and the 3 tests of vocabulary, 3 of the 5 C-Tests and the 2 holistic trans-

lation tests, all these tests may be unidimensional (see Table 6). What about other dimensions?

I don't believe that the 2 remaining C-Tests constitute a second dimension (see line 3). The 7 dictation parts are heterogenous. Perhaps different parts of the original text measure different aspects of language. Here a furhter analysis would be useful.

Finally, the MC-translation, a new method of measuring translation competence, does not fit. Perhaps this subtest constitutes a further dimension. But if we look only at the 11 tests which fit the model, we can conclude:

the 3 discrete point grammar tests
the 3 discrete point vocabulary tests
the 3 holistic C-Tests
the 2 holistic translation tests

are assumed to be one-dimensional and interval scaled. The name of this factor could be: general language factor in written language.

Bibliography

Kalveram, K.T. (1965), "Die Veränderung von Faktorenstrukturen durch simultane Oberlagerung". Arch. ges. Psychol. 117: 296-305.

Kalveram, K.T. (1970), "Problem der Selektion in der Faktorenanalyse". Arch. f. Psychol. 122: 199-230.

Moosbrugger, H. and Müller, H. (1982), "A Classical Latent Additive Test Model (CLA Model)". The German Journal of Psychology 6: 145-149.

Smith, V. and Klein-Braley, C. (1981), "Practical Approaches to the Teaching and Testing of Oral Skills for University Students in Germany". In: Klein-Braley, C. and Stevenson, D.K. (eds) (1981), Practice and problems in language testing. 1. Proceedings of the First International Language Testing Symposium of the IUS. Orbis Linguisticus 1. Frankfurt etc.: Verlag Peter D. Lang.

Smith, V. (1982), "On assessing speaking in small groups". In: Lutjeharms, M. and Culhane, T. (eds) (1982), Practice and problems in language testing 3. Proceedings of the Third International Language Testing Symposium of the IUS. Brüssel: Vrije Universiteit.

Thurstone, L.L. (1947), Multiple factor analysis. Chicago: University of Chicago Press.

FORMATIVE TESTS OF SPEAKING PROFICIENCY IN THE MOTHERTONGUE FOR 14-YEAR OLDS

Klaas Schreuder (Cito, Netherlands)

1 Introduction

For oral tests to be used in the course of the education process and meant to support it, that is for formative oral tests, as a rule only direct and not indirect testing methods will be adopted. That is to say, in taking such an oral test pupils will actually have to speak. And, indeed, I think that indirect measures of the oral skill at an intermediate level should be avoided, not only or in the first place for reasons of validity but with a view to

- the motivation of pupils, as well as
- the effects such tests have on education, whether intended or not.

Generally speaking testconstructors prefer their tests to elicit behaviour that does not differ too much from the criterion behaviour, the sort of behaviour that would occur in ordinary life and which is, in fact, what the test is meant to judge. In everyday reality there are, generally speaking, three situations calling for a demonstration of the oral skill: the monologue, the dialogue and the polylogue. In the large majority of situations falling into any of these three main categories, each speaker can, to a large extent, determine the direction of the activity in which he is engaged. This is selfevident in case of the monologue; in the other cases the influence exerted by the participants will not exactly be alike, but however that may be, it cannot be predicted with any certainty what the next speaker is going to say. Even situations that limit a speaker's free scope severely such as interviews, e.g. between a senior and a junior staff member, even such situations remain open-ended to a large extent. The junior staff member can cause the event in which he participates to take an unexpected turn. If it is thought important that the behaviour elicited by the test and the criterion behaviour do not diverge too much, the test should create a situation which leaves the testee the scope he would have in everyday reality.

Apart from this, 'realistic' tests will give rise to fewer methodological problems in validating tests, that is, provided the test's reliability meets the proper requirements. I take for granted that a test which is reliable and prompts behaviour that is quite close to criterion behaviour can be considered a valid test. Only if the behaviour required by a test deviates more markedly from the behaviour that is the real object of the evaluation, further validation is called for.

Of course I should make mention at this point of a difference between testing the oral skills of L1 and L2 learners. L1 learners can be assumed to avail themselves of a certain latitude in giving

direction to a conversation. But L2 learners at an elementary level will not regret the constraints put upon their freedom of expression: they would not know to use it if it were granted to them. That is why strictly structured tests of simple vocabulary, pronunciation or sentence structure are not at variance with the general principle of allowing the testee free scope. Incompatibility only arises at an advanced level of L2 proficiency. At this level requirements will be made of the L2 learner that come close to those made of the L1 learner. What I am going to say about oral tests of Dutch for 14 year old Dutch pupils might consequently be of interest to those concerned with testing L2 proficiency.

First of all I shall say one or two things about the starting points adopted for the construction of the tests. Then I shall discuss the test formats and present the results of pretests and rating try-outs.

2 Test content

The oral tests for native speakers of Dutch developed so far are part of a plan to compose a set of oral tests for the whole of secondary education. Tests for intermediate levels have been constructed first, because it is supposed that a series of such tests would contribute to the development of programmed instruction in the oral skill. That is to say, individual teachers as well as writers of educational material could benefit from the existence of a set of formative tests based on relevant educational objectives. An inventory of educational objectives through an analysis of most frequently used textbooks showed a lack of any sort of systematic approach.

The series of tests was set up as follows.

1 Oral proficiency was defined as: the skill to express oneself orally in an adequate manner in functional language situations. The latter are considered to be language situations derived from an analysis of requirements made of pupils both at school and out of school. The main criterion for categorizing these situations is the number of speakers involved: monologue, dialogue and polylogue. Each of the three main divisions is to comprise about ten language situations in accordance with the subcategories coming under the headings of monologue, dialogue and polylogue.

2 In selecting relevant subjects and situations we used the inventory presented in 'Basics in speaking and listening for High School Graduates' by Ronald E. Bassett et al. in Communication Education (1978: 293 f), which is an outcome of the minimal competencies movement in the USA. It lists all the situations calling for oral speech in which the average (American) citizen may get involved, situations that have to do with one's occupation, with citizenship and maintenace (private life). With a few adaptations to the Dutch situation, the list proved very helpful.

3 The tests' contents were largely determined by the relevance of subskills as apparent from their featuring in textbooks. Such subskills are:

reformulating, reasoning, distinguishing major matters from minor ones, ranking facts chronologically, expressing an opinion and (for the categories other than the monologue) asking and answering questions, contradicting, responding to the expression of emotions.
4 The way in which we wanted assessments to be made, was based on the following consideration. Since the tests are meant for classroom use enabling teachers to ascertain whether or not their pupils command a particular component of oral proficiency after instructions or without special training as the case may be-analytic assessment on the basis of a number of criteria will be more in the interest of the pupil and the teacher than global assessment. In this case agreement among raters would seem to be less important than consistency of the ratings over a period of time.

I shall now explain briefly with the help of some examples how the elements of this test content have been combined into tests. These elements are:

- speech situation (number of speakers)
- subject matter/context
- subskill.

A ('monologue') test puts the pupil in a situation in which he has to report to others on something that he has seen or heard. In this case it goes like this: a pupil is shown a ten minute film about the bio-industry. After some minutes' preparation he then has to report on the film to his class. How he goes about this is up to him. To aid him he does get a list of a number of obligatory points. Such a list proved an essential aid to memory for a great many pupils. The report is expected to be four minutes long. It is clear what we are after in this assignment: the pupil has to give a report (a subcategory of the monologue) in a situation which might occur at school or in a future job and in doing so he has to demonstrate the subskill of distinguishing between major and minor matters.

A second ('dialogue') test puts the pupil in a situation in which he has to carry on a conversation, not the informal everyday chat, but rather a more or less formal interview. The pupil is expected to acquire information by asking questions in some everyday situations. For instance, he is supposed to be wanting information about a periodical that he might subscribe to, about a sporting club that he is thinking of joining, about a particular organized trip that he might want to enter for. The pupil receives a sketch of the situation with the assignment to ask, within a five minute span, all the questions necessary to obtain the required information. Preparation time is ten minutes. After this there is the interview during which it is up to the teacher to supply the information asked for, but not all at once: on a number of occasions which the teacher selects himself he gives only evasive answers requiring the pupil to keep on asking questions until he has received a complete and clear answer. The pupil must not let himself be put off.

The subject of a third test, a 'polylogue', comes under 'citizenship', somewhat awkwardly perhaps, seeing that the test is meant for 13 to 14 year olds. Such a subject is 'public transport'.

Four pupils are given texts about public versus private means of transport, some days before the test is taken. Two of the pupils get a text which is biased in favour of private transport, the text the two other pupils get supports public transport. After studying their texts these four pupils have to carry out a particular assignment in discussing the subject. In principle the contributions of all four pupils are rated.

By means of these test formats we try to imitate as large a sample from real-life situations in which oral speech is used as possible.

3 Pre-tests and rating sessions

One of the constraints involved in the test situation is implied in the question whether the behaviour that is elicited can be assessed properly. To answer that question pretests and rating sessions were organized, the results of which I will discuss now. By means of the pretests we hoped to find out what was the best way of assessing the pupils' achievements. In the pretests the following tests were used: the monologue-test Report on a Film; the dialogue-test Asking Questions; the polylogue-test Discussion PUBLIC TRANSPORT.

The tests were set to pupils about 14 years old in second forms of secondary modern and lower vocational education. Each test was done by fifty pupils from various dialect areas all over the Netherlands. With a view to the rating sessions that would take place at a later date, the test sessions were taped: the monologue- and dialogue-tests on audiotape, the group discussions on video. Four rating sessions have taken place. The raters were teachers of Dutch who had some experience with the second form. Three of the four sessions will be discussed here.

These sessions were set up like this.

REPORTED RATING SESSIONS

No.	NUMBER OF RATERS	TEST	RATED CAND.	CRITERIA	SCALE POINTS	GLOBAL RATING
1.	11*	DIALOGUE* ANSWERING QUESTIONS	21*	11*	4*	X
2.	11*	DIALOGUE* ANSWERING QUESTIONS	21*	11*	4*	X
3.	12	MONOLOGUE REPORT OF FILM	27	4	4	X

*IDENTICAL

At the first two sessions, seperated by a week's time, the performances of 21 pupils on the dialogue test were assessed. At the third session the achievements of 27 pupils on the monologue test were evaluated. At all sessions analytic assessments on the basis of a set of criteria were made. Raters were also asked to give a global judgement on the basis of the usual Dutch ten-point scale. The following rating formats were used: an extensive model featuring eleven criteria and a four-point scale at the first two sessions, a more concise model with four criteria and a different four-point scale at the third session. The formats look like this.

RATING MODEL SESSION 1 and 2 DIALOGUE

CONTENT	SCALE POINTS YES	NO		
1. brings up the obligatory points	o	o		
2. brings up points of his own choosing	o	o		
	RARELY	SOME TIMES	OFTEN	ALMOST ALWAYS
3. asks for necessary elucidation	o	o	o	o
4. repeats himself unnecessarily	o	o	o	o
LANGUAGE				
5. words and sentences are suitable	o	o	o	o
6. pronunciation and words are non-standard	o	o	o	o
DELIVERY				
7. speaks clearly	o	o	o	o
8. speaks fluently	o	o	o	o
9. speaks monotonously	o	o	o	o
COMMUNICATION				
10. maintains contact with interlocutor	o	o	o	o
11. interrupts the interlocutor	o	o	o	o

GLOBAL RATING (marks from 1-10)

RATING MODEL SESSION 3 MONOLOGUE

	SCALE POINTS			
CRITERIA	WEAK	BARELY SUFFICIENT	FAIR	EXCELLENT
1. CONTENT	0	0	0	0
2. ORGANISATION	0	0	0	0
3. LANGUAGE	0	0	0	0
4. DELIVERY	0	0	0	0

GLOBAL RATING

3.1 Consistency and reliability of raters

The first question I mentioned above concerned the reliability and consistency of the raters. Consistency is, of course, an aspect of reliability. The consistency of raters could be computed as there were two sessions, a week apart, at which the same taped performances were assessed. We were particularly interested in this aspect because of the future use that would be made of the tests: as teachers will use them in class and as they will not be able to enlist the help of other raters, it is much more important for teachers to be consistent in their assessments than for teachers to agree with each other: consistency is more important than inter-rater agreement.

TABLE 1. COEFFICIENTS OF CONSISTENCY IN DIFFERENT RATING SESSIONS

RATER 1	.86
RATER 2	.85
RATER 3	.79
RATER 4	.84
RATER 5	.73
RATER 6	.73
RATER 7	.80
RATER 8	.73
RATER 9	.79
RATER 10	.74
RATER 11	.86

These correlations are relatively high which means that the raters are fairly consistent in their judgements.

In order to determine the rater's reliability - in the sense of their homogeneity - a homogeneity analysis was used. This procedure assigns numbers to scalepoints for each rater. These numbers are called scalevalues. By means of these scalevalues the original data (the rating sheets filled out by the raters) can be translated into a numerical datatable. For each rater the correlations between the perceived rating and the mean (= true) rating can be computed. The square of that correlation is the rater's reliability.
(Homogeneity analysis determines the scalevalues for the scalepoints in such a way, that the mean rater reliability - or homogeneity - is as high as possible.)
The rater's reliabilities for each of the three sessions were as follows.

TABLE 2. RATER RELIABILITIES-DISCRIMINATION MEASURES

RATER	RATING SESSION 1	2	RATER	RATING SESSION 3
1	.80	.81	1	.43
2	.74	.74	2	.67
3	.62	.69	3	.38
4	.71	.73	4	.58
5	.55	.59	5	.61
6	.50	.72	6	.43
7	.68	.67	7	.55
8	.63	.64	8	.43
9	.84	.76	9	.65
10	.63	.59	10	.51
11	.73	.75	11	.70
			12	.62
Mean	.68	.70		.55

The difference in mean reliability between the first two sessions and the third can perhaps be accounted for by the fact that different rating formats were used: the first format used eleven criteria and four scalepoints, the second format used four

criteria (and again four, but different scalepoints). The four criteria of the second format may be too complex to be applied unambiguously.
High mean reliability does not mean of course, that all raters agreed in their assessments. The analysis only determines the raters' homogeneity, that is, the extent to which their assessments pair, or also the extent to which one rater's assessment can be predicted on the basis of another rater's assessment. The raters will only give more or less similar assessments, if the scale values that are assigned to the scale points for each rater resemble each other per scale point, that is, if they have only little variance. The scale values obtained in the three sessions can be summarised as follows.

TABLE 3. SCALE VALUES. MEANS AND STANDARDDEVIATIONS.

RATING SESSION	SCALEPOINT	MEAN SCALE VALUE	STANDARD DEVIATION
1	RARELY	-.77	.07
	SOMETIMES	.00	.19
	OFTEN	.96	.19
	ALMOST ALWAYS	1.32	.18
2	RARELY	-.84	.09
	SOMETIMES	.11	.18
	OFTEN	.97	.16
	ALMOST ALWAYS	1.24	.19
3	WEAK	-1.21	.22
	BARELY SUFFI.	-.13	.14
	FAIR	.62	.11
	EXCELLENT	1.32	.26

From the differences between the mean values it can be concluded that the raters frequently used the same scalepoints, which means that they agreed in their assessments. For if this were not the case, if they had differed more or less randomly time and again, the mean scalevalues would have been identical.

From the fact that, in all three sessions, there is a regular increase of the mean value (even though the scale points used are not the same) it is clear that the scale points have been interpreted similarly to a certain extent, which is to say that, for example,

'often' has not systematically been regarded as superior to 'sometimes' in case of a particular criterion by some raters, whereas other raters dealt with this scale point the other way round. The differences between the mean values clearly show when reliability intervals are computed on the basis of the standard deviations found. As far as the column of standard deviation is concerned: the smaller the SD of the scalevalues assigned to a scalepoint is, the more frequently this scalepoint has been used by all raters at the same time. These standard deviations are low. Those for 'weak' and 'excellent' are highest for the third session. This is in accordance with the relatively low mean rater reliability for this session, as was shown in TABLE 2: these two scalepoints were apparently interpreted differently by different raters.

3.2 Independence of criteria

The second question regarded the independence of criteria. It was expected that the criteria could not be applied entirely independently, since they relate to one complex skill, but on the other hand they were not supposed to be connected too closely, as in that case far fewer criteria or even one criterion would suffice. And then the so-called analytic assessment would not have any advantages over a global one. Analytic assessment has the advantage, with classroom use of tests, that the pupils and their teacher can learn from the results at what points the pupils' skills fall short and call for additional training.

The homogeneity of the criteria has been determined by means of the same scale analysis was was used for computing rater reliability. For each of the three sessions a mean criterion was determined. If all criteria correlated high with this mean criterion, in fact one criterion would suffice. The next table shows the square correlations between the criteria and the mean criterion.

TABLE 4. DISCRIMINATION MEASURES CRITERIA

CRITERIA	RATING SESSION 1	2	CRITERIA	RATING SESSION 3
1	.07	.00	1	.62
2	.01	.01	2	.72
3	.25	.18	3	.59
4	.03	.07	4	.66
5	.68	.70	MEAN	.65
6	.41	.48		
7	.57	.67		
8	.57	.55		
9	.40	.57		
10	.57	.40		
11	.02	.00		
MEAN	.33	.33		

The mean correlations are low, particulary for the first two sessions. This means that the criteria do not overlap completely. This is mainly due to the criteria in the first category, that of CONTENT. Together with criterion no. 11 - 'interrupts interlocutor' - they are deviant. The four criteria used in the third session, which are as a matter of fact the headings of the four categories in the first format, all correlate more closely with the mean criterion. (These correlations cannot be simply compared to the others, because the criteria as well as the scale points differed from those used in the first two sessions.)

In order to establish to what extent the categories themselves are homogeneous, the correlations between the various criteria and the mean criterion per category have been computed as well. The next table shows a survey of the (square) correlations.

TABLE 5. DISCRIMINATION MEASURES PER CATEGORY OF CRITERIA

		SESSION 1	SESSION 2
CONTENT	1	.32	.00
	2	.26	.05
	3	.47	.66
	4	.24	.55
LANGUAGE	5	.78	.78
	6	.78	.73
DELIVERY	7	.71	.72
	8	.58	.63
	9	.58	.70
COMMUNICATION	10	.56	.57
	11	.56	.57

Once again, the criteria in the category of CONTENT appear to be deviant: they ill represent the category into which they fall.

The intercorrelations between the categories themselves can be clearly seen from the correlations between the mean criteria per category in the next table.

TABLE 6. CORRELATIONS BETWEEN CATEGORIES

RATING SESSION 1				RATING SESSION 2			
cat 2	.25			cat 2	.17		
cat 3	.27	.60		cat 3	.22	.66	
cat 4	.06	-.01	.06	cat 4	.13	.04	-.19
	cat 1	cat 2	cat 3		cat 1	cat 2	cat 3

Categories 2 LANGUAGE and 3 DELIVERY appear to correlate most closely. (Note that the minus signs are due to the peculiarities of the method for scale analysis used and can be ignored.)

From these data it appears that
- the CONTENT category should be split up;
- the criteria of the other categories represent the categories into which they fall fairly well;

- reducing the number of criteria to no more than the category headings is unnecessary;
- the categories of LANGUAGE and DELIVERY in the first format are the ones that could be combined.

3.3 Analytic and global rating

The last questions that we asked ourselves regarded the qualities of the two ways of assessment: global and analytic. The collected data do not point to a definitive conclusion. We preferred the analytic assessment, and this preference is not ruled out by the results. Here are some figures.

1 The correlation between the global marks obtained in the first two sessions is .69.
2 The correlations between the scores on the mean criterion and the global marks given by the raters are
session 1: .78
session 2: .55
session 3: .89
3 The first principal components of session 1 and session 2 (as far as the analytic assessment is concerned) show a correlation of .71. This value is to be seen as a test-retest reliability. From this it appears that
   - there is a close correlation between the two methods of assessment;
   - global marks would have sufficed for a ranking of pupils in the case of 11 raters;
   - for classroom use, when only one teacher can assess the achievements, the analytic assessment is to be preferred.

4 Bibliography

Bassett, R.E. et al. 1978. 'Basics in speaking and listening'. Communication Education 1978: 293 f.

Nishisato, S. 1980. Analysis of categorical data; dual scaling and its applications. Toronto: University of Toronto Press.

Rijlaarsdam, G.C.W. 1982. Beoordelen van discussievaardigheid. Amsterdam: SCO.

AN ALTERNATIVE MODEL FOR ASSESSING LANGUAGE PROFICIENCY AT THE END OF SECONDARY SCHOOL*

Elana Shohamy (Tel-Aviv University)

1. Introduction

Upon graduation from secondary school most students undergo some kind of evaluation of their achievement and proficiency. This evaluation is generally meant to assess students' achievements in relation to what they have been taught in school and/or to predict their performances in higher education. The evaluation often serves as the main criterion for acceptance and placement into higher education. It is a crucial assessment for the individual student since it can affect his/her future opportunities.

Educational systems around the world differ in the way they evaluate students' knowledge at the end of the secondary school. These differences can best be described on a continuum where on one side of it there are systems in which an external test is administred at the end of the secondary school, and the performance of the student on that test is his secondary school evaluation. On the other side of the continuum there are countries in which the school or the individual teacher assigns the end of secondary school evaluation without the intervention of any external body. In between these two extreme points there are systems in which some elements of the evaluation are done externally while others are done internally by the teachers in the schools.

It is difficult, and perhaps even irrelevant, to determine which testing system is preferable; the decision to adopt a certain testing system depends on a number of considerations which are specific to each educational system of a given country. In Israel, for example, a central system was adopted because at the time of the establishment of the country, when people immigrated from many different countries, there was a need for a central testing system which would serve as a unifying force for this diverse population. The external central test provided the Ministry of Education with an effective tool for carrying out this goal. Thus, till today, under the existing system students' knowledge in language and in other school subjects is assessed by the Ministry of Education via an external test administered at the end of secondary school. The national inspector of a given subject, with the aid of a small committee, write the test which is then administered to a select group of high school students (about 25 % of a given age level).

In the past few years there has been a call for giving schools more autonomy and independence and for possibly abolishing the external test at the end of secondary school.

The purpose of this paper is to describe a testing model which is currently being experimented with, and which could possibly serve as an alternative for the centralized testing system currently in use in Israel. In the next section a number of dimensions which are an integral part of any testing system will be presented, leading to a description of the proposed alternative model, its principles, assumptions and implementation procedures.

2. Testing dimensions

Zack (1981) lists a number of dimensions which are part of any testing system. The first testing dimension is the location of the testing body i.e. who determines what will be tested. This relates to whether the testing body is external such as a Ministry of Education, or internal such as the school or the class teacher. In centralized educational systems the assessment is generally done by the Ministry of Education while in more autonomous systems the evaluation is done internally by the students' teacher, or the school.

The second dimension is the time and frequency of the assessment. In some educational systems the assessment is done once, usually at the end of the twelfth grade. In other systems the evaluation at the end of secondary school is based on multiple assessments over a longer period of time. When the assessment is done once the student's score is determined by a single test; when the assessment is based on a number of evaluations it can also take into account his performance on projects, class participation and a number of tests. Thus, in a one-shot type of evaluation students' performance on the single test is crucial for their score - a failure or succes on that test is extremely meaningful. When the evaluation is continuous a student has a number of opportunities to perform and improve.

The third dimension mentioned by Zack is the number of students tested. In some systems only a small proportion of the students are being tested by the final secondary school test. This is related to the selection procedures which are generally based on the probability of a student to succeed on the final external test. In other systems all students of a given age level are being tested regardless of their probability to succeed on the test. When all the students of a given age are being tested, there is an increase in the opportunities of a larger number of students.

Another dimension of the end-of-secondary school evaluation is the purpose of the test - predictive, summative, or diagnostic. A test may predict students' knowledge in some given future situation, it may reflect the knowledge that was acquired in school without relating it to specific future situations, or it may provide feed-back and diagnostic information to teachers and students.

The last dimension which was not mentioned by Zack but nevertheless important, is the professionalization of the testing

body. This relates to the extent to which those who prepare the tests are qualified and trained in planning, writing and analyzing tests. Since the field of testing has become so sophisticated in recent years it requires qualification and training in areas such as measurement and psychometrics. Thus this dimension relates to the extent to which those who write the tests are professional testers. In centralized systems those who write the tests are often the subject supervisors who are knowledgeable in the area they supervise, but lack the appropriate knowledge in testing theory and practice. In educational systems where the teachers or principals prepare the test, they too lack this knowledge. Therefore, in some systems the task of writing the tests is given to a professional body such as a research institute or a testing agency.

The above represent a number of dimensions of the end of secondary school testing systems. It should be clear now that testing systems can differ by (a) the quality of the testing body - external or internal, (b) the frequency of the test administration - one shot versus a continuous evaluation; (c) the purpose of the test - predictive, summative or diagnostic; (d) the proportion of students tested - a select group or the total age group; and (d) the professionalization of the testing body - amateur, trained/amateur or professional. In the next section an alternative model will be described in the context of these dimensions.

3. The alternative model

The alternative model can best be understood by stating its four basic assumptions - pedagogical, professional, social and psychometric. These will be explained by way of comparing with the existing system. The assumptions will be related to the dimensions described above.

The **pedagogical** assumption of the alternative model is that the test matches the curriculum and not vice versa. This is contrary to the existing model where the test provides the tool for implementing the curriculum, and the teachers teach for the test.

The **professional** assumption is that the classroom teacher is a concerned individual who is capable and able of handling the secondary school evaluation task. His knowledge in testing procedures is obtained through extensive training by professional testers. As a result the teacher has the responsibility as well as the authority of assigning these evaluations. This is a different assumption than the existing system where the teacher is responsible for the testing of his students, but the authority of assigning the evaluation belongs to the external agency, the Ministry of Education.

The **social** assumption is that there is equal opportunity at the beginning of secondary school as well as at the end of secondary school. This implies that those who enter secondary

school receive an evaluation at the end and it is unrelated to their probability of performing on the external test as is the case in the existing system.

The psychometric assumption is that multiple tests and other forms of evaluation are administered to the students for determining their grades over a number of years. This leads to a more accurate and valid assessment, since the score is based on a larger sample of scores. Assigning a score based on one test, as is the case in the existing system, cannot be accurate enough psychometrically.

The above assumptions lead to the following operational principles of the alternative model as they relate to English as a Foreign Language (EFL):

1. Student achievements in EFL are assessed by their teachers who are trained by professional testers in testing procedures.

2. The assessment of students' knowledge in EFL is done continuously throughout the last three years of secondary school.

3. A national monitoring system provides comparative information for the schools and the entire educational system on the EFL proficiency of students.

Figure 1 displays the five dimensions of any testing system as they relate to the existing and the alternative testing systems.

Figure 1 DIMENSIONS OF TESTING SYSTEMS IN THE EXISTING MODEL AND IN ALTERNATIVE MODEL

Dimension	Existing system	Alternative
1. Location of testing body	External (Ministry of Education)	Internal (teachers)
2. Time	Once, at the end of 12th grade	Continuous - over a period of the last 3 years of secondary school
3. Purpose	Predictive, summative	Formative and summative
4. Proportion of students tested	Low	High
5. Professionalization of testing body	Amateur	Trained/amateurs

Dimensions 1 - 4 are taken from Zack, 1981.

It can be seen that in the first dimension - the location of the testing body, the existing system is external (the Ministry of Education) and in the alternative model, it is internal - the students' own teachers. In the second dimension the time, the existing system bases the evaluation on one test while the alternative model is based on a continuous evaluation over a period of 3 years. In the existing system the test's purpose is prediction and summary while in the alternative model it is formative. The continuous evaluation allows students to obtain feed-back and this is likely to affect their performance over the three-year period. In the third dimension under the existing system, only a small proportion of secondary school students reach the 12th grade since only those who are likely to pass the final test are selected to continue. In the alternative model, on the other hand, all students who attend secondary school, regardless of their likelihood to pass the external test, obtain the secondary school evaluation. In the last dimension, the professionalization of the testing body, in the existing model the testing body is amateur, while in the alternative model it is the students' teachers who obtain training in testing to become trained/amateurs.

4. Implementation of the alternative model

A description of the process by which the alternative model is implemented is given below.

4.1. Training of teachers

Since in the alternative model the EFL teacher is the one who assigns the evaluation at the end of secondary school and he is viewed as a professional, he is being trained extensively in the testing and evaluation of EFL achievement and proficiency by professional testers.

The professionalization is achieved in the following ways:

a) A number of workshops are conducted by the University testing team, and teachers learn about procedures and techniques for planning, constructing and analyzing language tests. Examples of topics which are included in the workshops are: test planning, writing, item and tasks, analysis of items and tests, testing the four skills, the use of rating scales, assigning grades, etc.

b) Monthly meetings are conducter in the schools between members of the university testing team and the individual teachers. The purpose of these meetings is to practice and implement the material learned in the workshops to the specific situations of individual teachers in their classess.

c) A tester's guidebook is distributed to the teacher, which contains a description of testing procedures, and ample samples of test items and tasks.

d) An item bank is constructed which includes a large number of test items and tasks which have been previously tried out and which contain psychometric data. This enables teachers to select items to suit their teaching objectives. Most of these items are written by trained teachers themselves, but have gone through statistical analyses by the testing team.

4.2. Following students' achievements over the last three years of secondary school:

Students performances on tests, projects and in class are recorded. Feed-back about this performance is constantly being fed to the students and to the teachers. Students' final score is generally an average of his performances in each of these activities each year.

4.3. A monitoring system

A national norming test is administrered in order to compare and adjust the scores of each test taker as assigned by the teachers. This ensures that the annual final score that the teacher has assigned is adjusted so it is not inflated or deflated by the teacher. Several procedures for adjusting the scores are experimented with, and tests in all the four language skills are developed. Thus, it is possible to obtain an estimation of the national EFL proficiency level and the performance of students in realation to this norm. A student's final score is a combination of the scores assigned by his teachers over a period of three years and his performance on this test. In the formula which combines these two components, the teacher's score counts more than the national test. The monitoring system is viewed as a temporary phase in the process of moving from the existing system to the alternative system. It is hoped that when teachers will become more professional as testers, there will not be a need for this norming system.

5. Conclusion

The alternative model is currently being implemented, on an experimental basis only, in a number of schools in Israel. It is now the fourth year of the experiment and there have already been students whose end-of-secondary school score was assigned by their teachers who had been extensively trained in testing. The students' scores were based on their performance in EFL over three years and the score that the teacher assigned had to be sometimes "adjusted" according to their class performance on the national monitoring system.

Several questions have come up during the research. Some can be solved and others may be viewed as deficiecies of the model. The questions relate to the extent to which teachers who already have "bad" testing habits can actually change their habits and start implementing new procedures. A further problem relates to the difficulty of accumulating scores over a long period and the weight that schould be given to students' achievement in each grade and for each area. A serious point of controversy is the monitoring system. Those who oppose it claim that if the assumption is that the teacher is a professional, why does tho score that he assigns need adjustment.

These are just some of the problems and questions raised. However, the advantages of giving more autonomy to teachers in testing as well as in teaching, the availability of teachers who know how to test, the existence of an item bank which can facilitate teachers' testing tasks, the continuous feedback to students and teachers, the fact that students' score is based on a large number of performances of different kinds, and the opportunity for more students to enter and complete secondary school education, all these seem to be more in line with the current mood and trend in the Israeli educational system.

Several research and evaluation projects are taking place within the experimental implementation of the model, in an attempt to judge its effectiveness in each of the three subjects (EFL, Mathematics and Hebrew writing). So far it seems that different conclusions are reached in each of these subjects.

However, it should be noted that the effectivenes of the alternative model must not be judged in comparison with the existing system, but rather by judging its own effectiveness and its feasibility for implementation. The decision of a country to use the alternative type of evaluation or the existing one depends on a number of factors. Does the country want to grant more authority to teachers, which means less authority to a central agency, which again means losing its controlling power? Does the country want a more democratic, more autonomous system and thus allow a larger proportion of students to enter and graduate from academic secondary schools? Does the country want to open its high school doors to all students? Does it view teachers as capable of taking the authority and responsibility? These and similar questions need to be answered before any new testing system is introduced. It is clear then that the testing system at the end of secondary school is a complex problem which involves, in addition to testing factors, educational, social, political, economical and philosophical considerations.

References:

Zack, Itai, et al. 1981. **Alternative evaluation for the matriculation examination, A workable Alternative.** Intermediary Report. The Unit for Measurement and Evaluation Research, Tel Aviv University.

* This article is written in memory of Professor Itai Zack whose belief in equal opportunities in education underlies the concept of the alternative model.
The alternative evaluation project is funded by the Israeli Ministry of Education and headed by Dr. Mordechai Miron.

ESTIMATING STUDENTS' VOCABULARY SIZES IN FOREIGN LANGUAGE TEACHING

Sauli Takala

1 Introduction

In this paper I will discuss some issues related to the estimation of people's vocabulary sizes and present some results from one large-scale assessment study. I will first outline different approaches to vocabulary research and then focus on the methodological problems related to quantitative estimation of acquired vocabularies. I will conclude by citing empirical results obtained from one study where some new ideas in test theory were applied to vocabulary learning.

2 Different approaches to vocabulary research

2.1 Why study vocabulary?

At the outset we should address the basic question: Why should anyone be interested in vocabulary research? Why should vocabulary knowledge be an interesting and important area for research? In sum, why bother about vocabulary? There are some indications that linguistics (e.g., Bolinger, 1963; 1970; 1976; Fillmore, 1979; Halle, Bresnan & Miller, 1978; Halliday 1966; Melchuk & Zolkovsky, 1974; Raskin, 1983) is showing a growing interest in the role of the lexicon and in lexical processes as an important part of linguistic theory. Psychologists and psycholinguists have demonstrated clearly for quite some time ago that vocabulary knowledge is the best predictor of reading comprehension (e.g., Anderson & Freebody, 1981). According to some estimates (e.g., Freebody & Anderson, 1981; Frumkina, 1967; Johnson, 1972; Klychnikova, 1973), about 70 % of the words in a text should be known for a global understanding of its meaning, about 90 % for understanding all main ideas, and about 95 % for understanding also details. Thus, we can conclude that vocabulary knowledge is definitely an important prerequisite for discourse comprehension, and seeing how central learning from text is in school and out-of-school, we have ample reason to maintain that vocabulary research is an important area for research and deserves, if anything, to be strengthened and intensified.

2.2 Approaches to vocabulary research

Vocabulary research can have a number of different approaches. In this paper I will discuss three such approaches. I will call them psychological, pedagogical, and quantitative, respectively.

If vocabulary research has a psychological bias, several questions arise as possible research problems. How is vocabulary processed in comparison to e.g., perception, syntax or whole discourse? What is meant by knowing a word? How does memory work in learning vocabulary (encoding, storage and retrieval) and how can different techniques (e.g., keyword method, hook method) possibly facilitate vocabulary learning? What causes difficulty and what facilitates vocabulary learning?

If vocabulary research has a pedagogical bias, several other questions merit attention. What words should be learned (issue of selection)? What should be the nature of learning outcomes at different stages of a course: beginning, intermediate, final stage (issue of objectives/goals concerning desired vocabulary knowledge and skills)? How should words be semanticized, i.e., how should their meanings be taught? How should word meanings be consolidated? What should be the role of conscious vs. incidental vocabulary learning?

If vocabulary research has a quantitative bias, as it may have due to its nature - consisting as it does of a large amount of different words - we may ask somewhat different questions. What is the total size of vocabulary in a language? How many different words do people know? How many words do ordinary people use, and how many words do writers use? How does vocabulary grow in childhood and in the later stages of life? How common are different words?

In order to get answers to such questions, several methodological problems have to be solved. What kind of test types can be used to test different kinds of vocabulary knowledge (validity issue)? How can we get good estimates of total vocabulary sizes on the basis of a sample of words (issue of research design, and problems related to reliability/dependability and generalizability)?

3 Estimation of students' vocabulary sizes

3.1 Problem

The main purpose of the study was to estimate the size of students' active and passive vocabulary in English after they had studied English for seven years (about 600 lessons, about 450 clock hours). For a more detailed description of the research problem, see the author's doctoral dissertation (Takala 1984).

3.2 Design

In this paper we are interested in estimating the overall size of English vocabulary learned by students in the Finnish comprehensive school. Thus we are dealing with program evaluation and domain-referenced (or criterion-referenced) measurement. We wish to generalize into the whole universe of content (i.e., taught vocabulary) and into the whole population of students. This means that it is necessary to specify the content domain and draw a random sample from it. Only this kind of design makes such two-way generalization possible. In such a design, it is useful or even almost necessary to apply multi-matrix sampling, which means that different students answer partly or totally different items. Thus several test forms are randomly rotated in class.

Population. The final target population of the study was defined as "all Finnish-speaking students in the final grade of 'normal' comprehensive school classes".

Student Sampling. Preliminary studies (Takala 1984) had shown that it is important to sample a sufficient number of schools, while it would not be necessary to sample many students from each school. The sampling method was a two-stage stratified cluster sample. The primary sampling unit was the school and the secondary sampling unit was the class. Four strata were used with the size of school and the degree of urbanization of the school community as the two bases of stratification.

The designed sample of school consisted of 42 schools and the executed sample of 39 schools. Altogether, 2,415 students took part in the study.

Item Sampling. Vocabulary size estimation promised to be a good starting point for generalizability studies. It is laborious but possible, due to Finland's fairly centralized school system, to define the domain and even list and count the items in the domain.

Two textbooks, which were practically the only ones used in schools, were reviewed and words taught in them were listed separately. Textbook 1 taught about 2,500 words for the two higher sets (Sets A and B) and about

1,500 words for the lowest set (Set C). Textbook 2 taught about 2,850 words and 2,340 words, respectively. From the two separate lists, a total of about 950 words was randomly drawn and distributed among 40 different test forms. Thus each student had to respond only to 40-50 items.

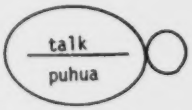
Certain design issues were tested in the study so that items were distributed to either "a robust student sample" and a "less robust student sample". They are not reported here (see Takala 1984).

3.3 Choice of test type

Several test types were considered. The constructed answer technique, in which students wrote the English equivalents of decontextualized Finnish words ("active vocabulary") and vice versa ("passive vocabulary"), was chosen on both theoretical and practical grounds. For a more detailed description of the rationale for the choice of the test type, see Takala (1984).

Sample items

Instructions: "In this test you can show how well you know the English vocabulary included in your course work. Below are presented a number of Finnish words. Your task is to write the English equivalent on the line above the Finnish word. Write the word even if you may not be quite sure about the correct spelling, since spelling mistakes are a minor consideration in scoring."



"Write the Finnish equivalents of the following English words."

4 Data collection and data analysis

Data on student vocabulary knowledge, and on the context of teaching and learning, were collected in the spring of 1979. Data file building took more than a year.

Student answers were scored 0-1 with meaning equivalence as the ultimate criterion (e.g., disregarding spelling). Interrater agreement was of the order of 95 %.

Data were analyzed using a logistic item analysis program and vocabulary size estimates were obtained through a new variance components analysis, which uses the generalized symmetrical sums (gss) method. It was shown that the results obtained with a new program are identical with those computed with Cronbach's formulas from the SPSS Reliability Program mean squares indices.

5 Some main results

The main results of the study can be briefly summarized as follows.

There was no reliable difference in the students' passive and active vocabulary knowledge, as they were measured in the study. Also, students' knowledge of simple word-formation rules and their contextual inference ability were poorly developed, in comparison to typical L1 skills. The following reasons were assumed: (1) Finnish and English are not related languages, which may not encourage such skills. (2) The emphasis at this stage is on syntactical patterns, while morphology is largely neglected. (3) The treatment of texts is "intensive", giving students little exposure to English. The estimated average size of vocabulary (see table 1, original estimates) was about 1,000 words, with great variability in performance. Fast learners knew about 1,500 words, average students about 900 and slow learners about 450 words. Due to the limited word-formation skills, the estimates ought to be adjusted by up to 45 %, by 17 %, and by 7 % for the three sets, respectively (see table 1, corrected estimates). The relationship between taught and learned vocabulary was 55 %, 32 %, and 20 % for the three sets, respectively.

Table 1. Original and Corrected Estimates for the Total Passive and Active Vocabulary Sizes, by Set

Set	Original estimates		Corrected estimates	
	Passive	Active	Active	Passive/content-aided
Set A	1,550	1,450	2,000	2,200
Set B	950	850	1,025	1,050
Set C	450	350	450	-

Variance components analysis showed that words made a greater difference in scores than students and that error of measurement can be lowered more efficiently by increasing the number of word items than by taking a larger student sample. There may also be an optimal size of input in vocabulary learning. Students who used a textbook with a lower input learned less than those whose textbook taught more words.

6 Implications and conclusions

Now that a new approach to a large-scale assessment of vocabulary size has been developed, tested empirically and found to be a promising line of study, several research questions suggest themselves. These can be divided into two major groups. One has to do with the test types and the other with student populations.

As was mentioned in the above, it was possible to test only limited aspects of vocabulary knowledge, namely relatively solid and easily accessible passive and active knowledge of words. Several experiments ought to be conducted with other test types that tap more partial knowledge of word meanings and see how vocabulary size estimates are effected.

Similarly, students' knowledge of vocabulary in the context of discourse comprehension and production ought to be estimated. Such experiments would provide data to complement the baseline data collected in the present study. It would then be possible to estimate, with a certain degree of confidence, that if students' decontextualized and firm knowledge of L2 words is X, their more partial knowledge of vocabulary is X + Y words, etc. It can be conjectured that partial knowledge of a fair amount of basic words combined with some knowledge of basic morphological rules and the availability of an adequate context can lead to an adequate comprehension of test passages and to provide a good opportunity for more word learning.

The study ought to be extended to other populations. With regard to the present study, it would be important to test students' knowledge of lower stage vocabulary at the end of that school stage. This would make it possible to explain with greater confidence the finding that lower stage vocabulary was known better than upper stage vocabulary. Is this so already at that stage or is lower stage vocabulary repeated during the upper stage, and thus the difference in learning is attributable to an increase in the opportunity to learn lower stage vocabulary? This question could be studied in even greater detail by looking at each successive grade and comparing the results.

Vocabulary size assessment should also be extended to older populations. How many words do students know at the end of the senior secondary school? How many words do L2 majors at the university know?

Other studies ought to address the question of how students' ability to use word analysis skills develops over time as the study of L2 progresses. Teaching experiments ought to be carried out in which students of different age levels are taught word analysis and context utilization skills in order to see what effect such direct teaching would have on students' vocabulary efficiency.

Further, since it was found that exposure to more words had a favorable influence on vocabulary learning, it should be studied what exposure leads to optimal word learning for students of varying ability. It seems likely that the relationship is not linear but more likely an inverted U-shaped curve.

In terms of curricular implications and educational equality concerns, it would be important to study when the observed large differences in vocabulary size in L2 emerge, and whether setting/streaming (and using different textbooks with different input) tends to increase or decrease such differences. Is limited input (i.e., smaller vocabulary size taught) better for slow learners or is that a misguided notion?

In addition to such empirical research, it would be useful to devote some attention to more theoretical questions on the nature of vocabulary learning, teaching, and research. Is it, for instance, in the very nature of a domain like vocabulary that the input should be large, and that the number of words known solidly would be low or conversely the number of words almost forgotten would be high? What would that mean for teaching, testing and grading? Is, for instance, the observed large item variance component an indication of the failure of teaching, or is it a natural characteristic of L2, and for that matter L1, learning and performance?

It is obvious that a whole research program is needed to increase our knowledge about vocabulary teaching and learning both in L1 and L2. Close links between L1 and L2 vocabulary research are of great importance for optimal progress. It may be more laborious to keep track of what is being done in both L1 and L2 research, but that is necessary to avoid duplication of effort and to utilize the state of art knowledge. This is one of the main lessons that work on this investigation has provided. It is time to put that belief into practice, now that the data invite further elaboration. This will be a rewarding experience, since vocabulary research tends to have a

special fascination of its own. Its range of interest is as wide as life itself. As Vygotsky so aptly put it, a word is a microcosm of human consciousness.

References

Anderson, R. C., and P. Freebody. 1981. Vocabulary Knowledge. In J. T. Guthrie (ed.) 1981. Comprehension and Teaching: Research Review. Newark, Del. International Reading Association. 77-117.

Bolinger, D. 1963. "The uniqueness of the word". Lingua 12: 113-136.

Bolinger, D. L. 1970. "Getting the words in". American Speech 45: 78-84.

Bolinger, D. 1976. "Meaning and memory". Forum Linguisticum 1: 1-14.

Fillmore, C. J. 1979. On fluency. In C. J. Fillmore, D. Kempler & W. S-Y. Wang (eds.). Individual differences in language ability and language behavior. New York: Academic Press. 85-101.

Freebody, P., and R. C. Anderson. 1981. Effects of differing proportions and locations of difficult vocabulary on text comprehension. (Tech. Rep. No. 202). Urbana: University of Illinois at Urbana-Champaign, Center for the Study of Reading.

Frumkina, R. M. 1967. "Slovar-minimum i ponimanie teksta" (Minimum vocabulary and text comprehension). Russkij jazyk za rubezom 2: 15-21.

Halle, M., Bresnan, J., and G. A. Miller. 1978. Linguistic Theory and Psychological Reality. Cambridge, Mass.: The MIT Press.

Halliday, M. A. K. 1966. Lexis as a linguistic level. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, and R. H. Robbins (eds.) 1966. In memory of J. R. Firth. London: Longmans. 148-162.

Johnson, D. B. 1972. "Computer frequency control of vocabulary in language learning reading materials". Instruction and Science 1: 121-131.

Klychnikova, Z. I. 1973. "Psychologiceskie osobennosti obucenija cteniju na inostrannom jazike" (Psychological peculiarities of teaching reading in a foreign language). Moscow: Nauka.

Melcuk, I. A., and Zolkovsky, A. K. 1974. Towards a functioning 'meaning-text' model of language. In V. Ju. Rozencvejk (ed.) 1974. Essays on lexical semantics. Vol. 2. Stockholm: Språkförlaget Skriptor. 1-53.

Raskin, V. 1983. A concise history of linguistic semantics: I. West Lafayette: Purdue University.

Takala, S. 1974. Evaluation of students' knowledge of English Vocabulary in the Finnish Comprehensive school. (Tech. Rep. No. 350). University of Jyväskylä. Institute for Educational Research.

EVALUATING THE LANGUAGE LABORATORY IN PRACTICE

Robert Vanderplank (Helsinki University)

1 A model for language laboratory evaluation

1.1 Background

Some time ago I investigated the proposition that the language laboratory (henceforth, LL) was largely a wasted resource. I wanted to find out why a valuable educational tool, written about for over 25 years, often at great length, and in use all over the world, was so poorly regarded by many and rarely seemed to come up to expectations when put to the test. This work was greeted well by some language schools in Britain, but on the whole it was received with an embarrassed silence. I had hoped to raise standards of LL use by attempting to raise user-consciousness through the evaluation techniques which will be described below.

The literature on LL evaluation is rather depressing. Most studies, for example, the Pennsylvania Project (Smith, 1970), the 'communicative competence' study of Savignon (1972), and the York Study of Green and his associates (Green, 1975) come out against the LL. A notable early exception is the study of Sarah Lorge (1964), but on the whole that study remains an exception. Indeed, the last large-scale study in Britain, the York Study mentioned above, was particularly damning: that current uses of the LL make it largely a waste of money.

On the other side, there is a great deal of literature on how to select a LL, how to manage it, how best to exploit it, its advantages, and so on. Then there is a third branch which deals with reactions of teachers: questionnaires, attitude surveys, such as those by Anderson (1977) in Sweden, and by Holec (1971) in Belgium.

When, early on in my research, I complained about the basic unfairness of LL comparative studies, since they were loaded against the LL from the outset by their very design, I was told that there was no other way. As a convinced LL man who has worked in LL-orientated environments and is conscious of the benefits it may bring, I could not accept the findings of many studies in any way other than trivial. If you reduce your expensive tool to an almost insignificant role, in order to compare its performance with some thing else, waste much of its potential and fail to train personnel and make them aware of its potential, then it seems to me that you are not holding a fair evaluation.

1.2 How the model is made up

I wanted to find a means of evaluating the LL on its own terms in practice. That is, how this sophisticated, educational tool, with its own rationale and principles of good use, and with sound pedagogical reasons to justify its purchase, was actually used by teachers and learners (leaving aside purely administrative reasons for its use). I

took as my starting point the three basic reasons for the purchase of any new tool: that it does the job FASTER, BETTER, or gives more USER SATISFACTION than any other comparable piece of equipment or means of doing the job. This would mean, in teaching terms, the tape recorder, a simple form of LL, such as an audio-active LL, or no educational aids.

How are we to translate these criteria into LL terms? Well, writers on LL use often talk about the advantages of using the LL. I looked at the literature and found some 13 given, but there are seven which all writers agree on:

1. Each learner can answer all the questions and work all the time.
2. Each learner is responsible for his own performance.
3. Each learner can listen critically to his own voice.
4. Each learner can work at his own pace.
5. The teacher can deal with each learner's problems individually.
6. The LL can provide a variety of programmes and activities.
7. Learners are not afraid to speak in the privacy of the LL.

How can we link the criteria of FASTER, BETTER, and JOB SATISFACTION with the proposed or assumed advantages given in the literature. I would suggest the following:

WORK STUDY CRITERIA	ASSUMED ADVANTAGES
1. Quickness in performing tasks (cutting time/increasing practice time per student)	own pace, answer all questions
2. Accuracy in performing tasks (cutting waste/improving class performance)	listen critically, teacher can help individually
3. Job Satisfaction (using the intelligence of the user as a contributing factor to the above two criteria/ increasing interest and motivation)	responsibility, privacy, variety

You may well suggest that in reality the links are not quite as neat and clear-cut as I have made them. I would agree entirely, and this overlapping will be accomodated in another part of the model which will be described below.

What happens in practice? Of course, the advantages do not come automatically, although LL salesmen might suggest that they do. Meeting the criteria and gaining the advantages depend on whether specific facilities built into the LL are used and also on how they are used. In other words, the advantages are only realised in practice by the use of the facilities available and by the fulfilment of specific conditions regarding the use of the facilities. Let us take, for example, assumed advantage number 3. Each learner can listen critically to his own voice. In strictly practical terms, for this assumed advantage to be obtained, not only must the LL have a record/playback facility and dual-track tape recorder, but the learner must also be able to manipulate controls, drills and exercises with ease, independent

of the teacher, and have both time and the ability to assess what has been recorded and then correct errors. Moreover, the learner must be aware that the responsibility for the performance is his alone (assumed advantage number 2.).

In more general terms, each assumed advantage can be said to be based on the presence and use of facilities (e.g. cue/rewind, monitor/intercom.) or functions (learner can recap/correct, teacher can listen to learner performance/correct learner), and also on the fulfilment of both pedagogical and technical conditions. The facilities and functions of a normal AAC LL are well-known and I shall not go through them here. As far as pedagogical and technical conditions are concerned, my model includes some thirty in all, some of which are shown below:

Pedagogical conditions of use

LL work should be purpose-designed
Teacher should be trained in exploitation of LL
Objectives of any LL session should be explicit
Amount of LL work should permit learners to work systematically through at own pace
Material should be pre-recorded if administratively possible
Learners must be trained to avoid over- and under-learning
Learners must be trained not to rely on teacher intervention
Students must be trained in use of LL controls
Material used must be adequate and appropriate for the task

Technical conditions of use

Equipment used should be adequate for the task in terms of noise, reliability, clarity
All facilities should be functioning
Material used should be technically adequate e.g. quality of recording

I think it should be clear by this point that some facilities and functions and their underlying conditions of use can be linked to almost all assumed advantages, while others are more specific. For example, those concerned with self-assessment and correction, and possibly with over- and under-learning can be linked directly with assumed advantage number 3. Each learner can listen critically to his own voice. If we take this point a stage further, we can say that some facilities/functions and conditions have greater bearing on one of the three criteria for LL use: SPEED (S), ACCURACY (A) and JOB SATISFACTION (J), than on the other two. That is, you can weight the use of a facility or an underlying condition with S, A, or J, or with any combination of them. The way this works is shown below in the Operational Model for the analysis and evaluation of LL use.

Let me summarise briefly what I have said so far. My model is built on :

1. utilization of facilities and functions, and,
2. fulfillment of conditions underlying their use.

The absence and presence of these conditions and the degree to which any facility or function is exploited act as deciding factors in whether an assumed advantage is obtained in practice in any LL session.

OPERATIONAL MODEL FOR ANALYSIS AND EVALUATION OF LL USE

CRITERION	ADVANTAGE	USE OF FACILITY/FUNCTION or CONDITION OF USE	PRACTICE WEIGHTING
SPEED	1. ANSWER ALL Q'S & WORK ALL TIME	LL WORK PURPOSE-DESIGNED	S/A
		START-UP TIME SHORT	S
		LEARNERS PARTICIPATE IN SESSION	J
		LEARNERS' INTEREST IN SESSION	J
		TEACHER 'HAPPY' USE OF LL	S/A/J
		TEACHER LIKES USING LL	J
		TEACHER EXPERIENCED	S/A
		LL WORK COORDINATED	J
		MATERIAL INDEXED	S
JOB SATISFACTION	2. RESPONSIBLE FOR OWN PERFORMANCE	ALL LEARNERS CLEAR ABOUT OBJECTIVES	S/A/J
		ALL L. 'HAPPY' MANIPULATING DRILLS	S/A/J
		FREQUENT USE OF 'T' CALL	J
		DISCRIMINATING USE OF 'T' CALL	S/A
		PEDAGOGICAL MONITOR	A/J
		INSTRUCTIONAL MONITOR	S/J
		ANALYSIS/DIAGNOSIS OF PRON. ERRORS	S/A/J
		ANAL/DIAG. OF STRUCTURAL ERRORS	S/A/J
ACCURACY	3. LISTEN CRITICALLY TO OWN VOICE	ALL LEARNERS ABLE TO CRITICISE	A/J
		ALL CAPABLE OF SELF-CRITICISM	A
		LEARNERS TEND TO UNDER-LEARN	A
		LEARNERS TEND TO OVER-LEARN	S

CRITERION	ADVANTAGE	USE OF FACILITY/FUNCTION or CONDITION OF USE	PRACTICE WEIGHTING
SPEED	4. WORK AT OWN PACE	ALL LEARNERS 'HAPPY' MECHANICALLY	S/J
		ALL PRE-RECORDED	S/J
		LEARNERS CONTROL DURING SESSION	J
		LEARNERS ABLE TO VARY PACE	S/J
		LEARNERS 'HAPPY' WITH PRE-SET PACE	J
		VARIETY OF PACE	J
ACCURACY	5. TEACHER CAN DEAL WITH INDIVIDUAL LEARNERS	TEACHER MONITORS LEARNERS IN SESSION	A/J
		INDIVIDUAL/GENERAL MONITOR	S/J
JOB SATISFACTION	6. VARIETY OF PROGS. & ACTIVITIES	VARIETY OF MATERIALS	J
		VARIETY OF ACTIVITIES	J
		LL FUNCTIONS EXPLOITED	J
		RECORDED MATERIALS INTEGRATED	S/J
JOB SATISFACTION	7. PRIVACY OF BOOTH	ACOUSTICS/NOISE	J

In turn, the absence or presence of the different advantages then indicates whether the LL session can be said to have been held for reasons of SPEED, ACCURACY, or JOB SATISFACTION, none of these, or any combination of them.

2 The model in practice

2.1 Observation grids

The model was adapted to a set of observation grids, four in all. These are shown overleaf. The most important grids for assessing individual LL sessions were Grid 2. Teacher use of LL and Grid 3. Student use of LL .

As can be seen from grids 2 and 3, absence or presence of conditions and use of facilities and functions was not enough. I was also interested in the degree to which equipment was used and functions were carried out. Each scale, therefore, had a set of operational definitions. Some examples of the definitions used are given after the grids. '0' - zero - was always taken to mean 'inapplicable in this session'.

It may all seem rather 'ad hoc' to the informed reader. I should say at this point that I attempted to gain validity for the grids and definitions (which were, in fact, derived entirely from the literature and findings of research on LL use) by distributing questionnaires on LL use to all teachers in the first school studied. The results of these questionnaires supported very strongly the model, the content of the grids and the definitions. It could be claimed, therefore, that teachers were being assessed by the very criteria that they themselves accepted or supported.

2.2 The observations

There were seven observers in four schools of English (including the author). The schools were the School of English Studies, Folkestone, Colchester and Bedford English Study Centres, and the Davies School, Cambridge. Unfortunately, the teacher at the Davies School who was to carry out the observations proved to have neither the knowledge or the experience to carry out the task adquately. It should be clear by this stage that operating this model with any degree of accuracy and reliability does require a sound knowledge and understanding of LL's and considerable experience with them. Thus, even a disappointing observer was, for the purposes of the trial of the model, useful in some respects.

Altogether 56 LL sessions were observed. In each session the observer completed the marking of the grids as the session progressed. Some class information regarding materials, students and the teacher was, of course, known in advance, so grids 2 and 3 could receive most attention. By the end of each session, an observer had gathered a lot of quantitative data on the use of facilities and functions and on the extent to which conditions had been fulfilled. These data could then be grouped according to their bearing on each assumed advantage, using the weighting given to each in terms of S, A and J.

GRID 1 : CLASS DESCRIPTION of class ..... with ..... at .... on .

Learning:	Content of course:
	Objectives:
	Time:
	Frequency:
	Learning environment:
	Material:
	Progressive rigour:
Students:	Socio-cultural:
	Intellect:
	Aptitude: 1. for learning:
	2. for language learning:
	Motivation:
	Level of L2:
Teacher:	Attitude to students:
	Attitude to method:
	Attitude to LL:
	Length of experience with LL:
	Cultural awareness:
	Linguistic and psycholinguistic knowledge:
Class:	Number of students:
	Average age: max. age: min. age:
	Class level:
	L 1's
	Use of LL before:

Teacher native speaker/non-native speaker

Teacher's age:

GRID 2: OBSERVATION OF CLASS ..... with ..... at ..... on .....

LAB. SESSION: circle appropriate alternatives

integrated/supplementary
class mode/library mode
supportive/initial
tape only/multi-media/realia/printed matter
cognitive/manipulative

TEACHER'S USE OF LL DURING SESSION: Circle appropriate number using definition sheet.

T. 'happy' use of LL	5 4 3 2 1 0
T. likes using LL	5 4 3 2 1 0
T. experiences in LL use	5 4 3 2 1 0
Objectives of session explicit	5 4 3 2 1 0
Start-up time short	5 4 3 2 1 0
Material indexed	5 4 3 2 1 0
LL work co-ordinated with class work	5 4 3 2 1 0
LL work purpose-designed	5 4 3 2 1 0
LL functions exploited	5 4 3 2 1 0
All pre-recorded	5 4 3 2 1 0
Variety of materials	5 4 3 2 1 0
Variety of activities	5 4 3 2 1 0
Variety of pace	5 4 3 2 1 0
Choice of materials	5 4 3 2 1 0
Choice of pace	5 4 3 2 1 0
T. control of pauses	5 4 3 2 1 0
Recorded materials integrated with other materials in LL	5 4 3 2 1 0
teacher monitors students/time	5 4 3 2 1 0
Pedagogical monitoring/time	5 4 3 2 1 0
Instructional monitoring/time	5 4 3 2 1 0
Monitoring from individual to general	5 4 3 2 1 0
Teacher makes notes of student errors	5 4 3 2 1 0
Analysis and diagnosis of student pronunciation errors	5 4 3 2 1 0
Analysis and diagnosis of student structural errors	5 4 3 2 1 0
Teacher feed-in of supplementary materials	5 4 3 2 1 0

BEST COPY AVAILABLE

GRID 3: OBSERVATION OF CLASS ..... with ..... at ..... on .....

STUDENTS' USE OF LL DURING SESSION: Circle appropriate number using definition sheet.

		COMMENTS
All students clear as to pedagogical function of LL	5 4 3 2 1 0	
All S. 'happy' manipulatively	5 4 3 2 1 0	
All S. 'happy' mechanically	5 4 3 2 1 0	
S. tend to 'overlearn'	5 4 3 2 1 0	
S. tend to 'underlearn'	5 4 3 2 1 0	
All S. clear as to objectives of session	5 4 3 2 1 0	
S. able to vary pace	5 4 3 2 1 0	
S. 'happy' with pre-set pace	5 4 3 2 1 0	
S. able to self-correct/evaluate	5 4 3 2 1 0	
S. capable of self-correcting/evaluating	5 4 3 2 1 0	
S. frequent use of T. call	5 4 3 2 1 0	
S. discriminating in use of T. call	5 4 3 2 1 0	
S. control 100%/time	5 4 3 2 1 0	
S. interested/time	5 4 3 2 1 0	
S. participating/time	5 4 3 2 1 0	
S. individualised learning/time	5 4 3 2 1 0	

HARDWARE USED IN LL SESSION

Response requirement adequate for task:

LL system: yes/no
microphone: yes/no
headset: yes/no

T. has full remote control: yes/no

'S' stop when T. intercom.: yes/no

Functions available: high-speed transfer/simultaneous S. & T. record/random access revolution counter/loop-repeater/multi-input/multi-programme/two speeds/large reels for spools/pause control/'S' record indicator/writing surfaces/group & conference facilities /

Functions used in LL session:

LL layout: Wall facing/rows with sightlines/rows without sightlines

Noise: yes/no acoustics: good/OK/poor dust: yes/no

Availability of other equipment: easy/difficult

APPENDIX I.22. GRID 4 : SECOND STUDY

GRID 4: OBSERVATION OF CLASS ..... with ..... at ..... on

MATERIALS USED IN LL SESSION: Circle appropriate choice(s).

COMMENTS

Variety of material available at this level: yes/no
Prepared for LL use/prepared for textbook/multi-purpose
Published material/in-house material/scripts/teaching notes

Production factors:

clarity	good/adequate/poor
intonation	good/adequate/poor
tone	good/adequate/poor
speed	good/adequate/poor
noise	good/adequate/poor
accent	good/adequate/poor
intervals	good/adequate/poor

Recorded materials: 1. exercises/drills

Compatible with other materials	yes/no/unclear
Unambiguous instructions	yes/no/unclear
Str. content controlled/graded	yes/no/unclear
Changes from segment to segment controlled/graded	yes/no/unclear
Sufficient examples of str. pattern before change of pattern	yes/no/unclear
Single correct response	yes/no/unclear
Vocabulary selected/graded	yes/no/unclear
Sounds well-related to material accompanying	yes/no/unclear
Variety of drill type	yes/no/unclear

2. dialogues for imitation/memorization/role-playing

Authentic cultural setting	yes/no/unclear
Natural subject matter	yes/no/unclear
Useful phrases for memorization	yes/no/unclear
Segments for repetition of 'memorable' length	yes/no/unclear
Background noise appropriate	yes/no/unclear

3. other materials used

Is material used at other levels/in other situations	yes/no/unclear
Is hardware present for full exploitation	yes/no/unclear
Is content/time ratio adequate	yes/no/unclear

Differential index: Circle appropriate number on each scale using definition sheet.

discrete	+ 3 2 1 0 1 2 3 -	integrated
inflexible	3 2 1 0 1 2 3	flexible
reactive	3 2 1 0 1 2 3	interactive
linear-systematic	3 2 1 0 1 2 3	indexed-selective
T. controlled	3 2 1 0 1 2 3	S. controlled
imitative	3 2 1 0 1 2 3	creative
phonetic	3 2 1 0 1 2 3	semantic
structural	3 2 1 0 1 2 3	functional
devised/adapted	3 2 1 0 1 2 3	authentic
skill-based	3 2 1 0 1 2 3	concept-based
habit-based	3 2 1 0 1 2 3	memory-based

Will materials used help meeting teaching objectives in terms of skills, content, grading ? yes/no/unclear

BEST COPY AVAILABLE

DEFINITIONS

TEACHER 'HAPPY' USE OF LL:

5 - TEACHER MAY DISLIKE USING LL OR MAY BE INEXPERIENCED IN ITS USE BUT DEMONSTRATES A MANIPULATIVE AND MECHANICAL SKILL WITH CONTROLS AND FUNCTIONS WHICH INDICATE AN AWARENESS NOT ONLY OF THE MECHANICAL OPERATION OF THE EQUIPMENT BUT ALSO OF ITS INHERENT METHODOLOGICAL ADVANTAGES AND LIMITATIONS.

3/4 - NEUTRAL PERFORMANCE, i.e. NO PARTICULAR SYMPATHY BUT APPARENTLY AWARE OF THE NATURE OF THE EQUIPMENT AND APPROPRIATE MANNER OF EXPLOITING IT.

1/2 - TEACHER MAY LIKE USING THE LL AND MAY EVEN BE EXPERIENCED IN ITS USE, BUT DEMONSTRATES AN ABSENCE OF MANIPULATIVE AND MECHANICAL SYMPATHY WITH CONTROLS AND FUNCTIONS WHICH INDICATES A LACK OF AWARENESS NOT ONLY OF THE NATURE OF THE EQUIPMENT, BUT ALSO OF ITS INHERENT METHODOLOGICAL ADVANTAGES AND LIMITATIONS.

ALL LEARNERS 'HAPPY' MANIPULATIVELY:

5 - LEARNERS ABLE TO PERFORM THE STANDARD VARIETY OF DRILLS AND EXERCISES, e.g. REPETITION, TRANSFORMATION, SUBSTITUTION, etc., AND TO MOVE FROM ACTIVITY TO ACTIVITY WITH MINIMUM LOSS OF CONCENTARTION.

3 - SOME LEARNERS ARE NOT SO 'HAPPY' BUT ARE ABLE TO BENEFIT FROM PRACTICE.

1 - LEARNERS DEFINITELY 'UNHAPPY' UNCLEAR AS TO HOW TO PERFORM DRILLS (c.f. TEACHER INSTRUCTIONAL MONITOR/TIME)

INSTRUCTIONAL MONITORING/TIME:

5 - 100% of MONITORING TIME

4 - 75%

3 - 50%

2 - 25%

1 - 0%

DEFINITIONS

TEACHER EXPERIENCED IN LL USE:

FREQUENCY X PER WEEK	DURATION FRACTION OF HOUR	X	LENGTH OF TIME MONTHS

5 - 100+ VERY EXPERIENCED

4 - 80+ MODERATELY EXPERIENCED

3 - 60+ ADEQUATE EXPERIENCE

2 - 40+ LIMITED EXPERIENCE

1 - 20+ INADEQUATE/INEXPERIENCED

LEARNERS TEND TO UNDER-LEARN

5 - OVER 50% OF LEARNERS HAVE TENDENCY TO RUSH THROUGH EXERCISE AND DRILLS WITHOUT SUFFICIENT ATTENTION TO EVALUATION OF PERFORMANCE AND WITH INSUFFICIENT USE OF PAUSE AND REPLAY FACILITIES

3 - SOME LEARNERS; BUT UNDER 50%, UNDER-LEARN AT SOME POINTS IN THE SESSION

1 - FEW, IF ANY, LEARNERS TEND TO UNDER-LEARN AT ANY TIME

LEARNERS ABLE TO VARY PACE

5 - LEARNERS HAVE COMPLETE CONTROL OF PACE OF LEARNING AND PRACTISING THROUGHOUT SESSION

3 - LEARNERS HAVE CONTROL FOR OVER 50% OF SESSION e.g. AFTER TRANSFER OF PROGRAMME IN LOCK-STEP

1 - LEARNERS WORK IN LOCK-STEP THROUGHOUT SESSION

BEST COPY AVAILABLE

As was said earlier, some conditions or use of facilities/functions have a bearing on all three criteria and so were given equal S,A and J weighting, but at the same time were felt to be more relevant to a single assumed advantage and consequently were attached to that advantage. Since assumed advantages were also given S, A or J criterion, obtaining an assumed advantage establishes whether the respective criterion is being fulfilled, and thus, a qualitative assessment in terms of Speed, Accuracy or Job Satisfaction (i.e. their pedagogical equivalents) is then possible.

The results for eleven out of the fifty-six sessions are shown below (a representative sample). Two of these sessions, 19 and 39, are also given the analysis described above. The analysis shows that the main criterion for holding session 19 appears to have been Job Satisfaction. Of course, in pedagogical terms, without the other criteria, it could well be interpreted as simply giving the learners a change of scene and activity, with no clear pedagogical objectives. In session 39, on the other hand, the Job Satisfaction criterion is linked to the Accuracy criterion, through an emphasis on both accuracy and self-responsibility.

The full results in terms of the three criteria were as follows:

Criteria	Number of sessions
Job Satisfaction/Speed/Accuracy	13
Job Satisfaction/Speed	13
Job Satisfaction/Accuracy	7
Job Satisfaction only	13
Speed/Accuracy	1
Accuracy only	0
Speed only	5
None	3
Insufficient data	1

Only thirteen out of fifty-six can be said to have exploited the facilities of the LL fully according to the criteria, and to have gained the advantages of LL use. So, just as the York Study found, the LL really is under-exploited and in these terms is a wasted resource for many. The reasons for the poor showing of so many sessions have been discussed elsewhere (Vanderplank, 1981), but limitations of space means that I can only briefly summarise what the good sessions had in common and what the limiting factors in LL use appear to be from this study.

3 Common features of fully-exploited sessions

The thirteen fully-exploited sessions had the following common features:

- frequency of use (once a day or more)
- teachers skilled in LL use
- students well-trained and responsible (all machines fully under their control).

Assumed advantage	Grid conditions	S/A/J	SESSIONS 1	2	3	16	17	18	19	20	37	38	39
Each student can answer all q's and work all the time	LL work coordinated		3	5	3	2	4	3	3	2	3	3	5
	T 'happy' use of LL	S/A/J	3	3	5	3	5	3	3	5	4	3	3
	T likes using LL	/J	4	2	2	3	5	4	4	5	4	5	5
	T experience	S/A	5	5	5	3	5	3	5	5	5	5	5
	Mat. indexed	S/	5	4	5	5	5	5	5	2	5	5	0
	Start up time short	S/	4	5	5	4	5	3	5	4	5	5	5
	S interest/time	/J	5	4	5	3	5	5	5	5	5	4	5
	S participation/time	/J	5	4	5	4	5	5	5	5	5	4	5
	LL work purpose-designed	S/A	5	5	3	3	4	2	3	4	4	3	5
Each S. responsible for own performance	All S happy manip.	S/A/J	4	2	3	3	4	4	3	5	4	3	5
	All S. clear objectives	S/A/J	4	5	4	3	4	5	4	4	4	4	5
	Frequent T-call	J	2	2	2	2	1	2	2	2	2	1	1
	Discrim. T-call	S/A/J	1	2	2	1	0	1	3	5	5	0	0
	Pedagogical monitor	A/J	2	2	4	1	2	3	2	1	2	4	5
	Anal/diag. pron. errors	S/A/J	4	3	3	2	3	3	3	0	1	3	5
	Anal/diag. str. errors	S/A/J	2	3	3	2	1	2	1	0	1	3	5
	Instructional monitor	S/ /J	4	4	2	5	2	3	4	3	4	2	1
S. can listen critically to own voice	All S. able crit.	A/J	2	5	4	4	4	3	4	4	3	3	5
	All S. capable crit.	A/	2	2	2	2	4	3	3	0	2	2	4
	S. tend underlearn	A/	3	3	3	4	3	4	4	1	3	3	1
	S. tend overlearn	S/	1	1	1	1	1	1	1	1	1	1	1
S. can work at own pace	All S. 'happy' mech.	S/ /J	5	4	4	2	4	3	4	5	4	5	5
	All pre-recorded	S/ /J	1	1	1	1	1	1	5	1	1	1	5
	S. control/time	J	3	3	4	4	3	3	4	4	3	3	5
	S. able to vary pace	S/ /J	1	1	0	3	4	3	4	3	4	2	0
	S. 'happy' pre-set pace	/J	4	3	3	3	3	4	0	4	2	1	1
	Variety pace	/J	1	2	1	1	3	1	1	2	4	5	5
T. can deal with individual students	T. monitor S./time	A/J	3	5	4	4	4	4	4	2	1	1	1
	Individ-general monitor	S/ /J	2	1	1	1	1	1	1	1	3	3	3
LL can provide variety of programmes & activities	Rec. mats. integrated	S/ /J	1	2	3	2	4	3	3	3	3	3	1
	Variety of materials	/J	2	1	2	1	3	1	2	2	2	2	1
	Variety of activities	/J	2	1	2	1	3	1	2	2	3	3	3
	LL functions exploited	/J	4	3	3	3	4	3	3	3	✓	✓	✓
S. can work in privacy of booth	Acoustics/noise	/J	✓	✓	✓	✓	✓	✓	✓	✓			

SESSION 19

ADVANTAGE 1. J and S weightings predominate. Advantage obtained.

ADVANTAGE 2. Mixed weightings. The advantage is not fully obtained as the teacher must spend so much time telling some learners what to do.

ADVANTAGE 3. Many of the class are unable to take advantage of the compare facility.

ADVANTAGE 4. The advantage gained by pre-recording is largely cancelled by a monotonous pace of work throughout the session, with only a short song to break it at the end. High motivation shown by high 'interest' score.

ADVANTAGE 5. C.f. Advantage 2. Most communication with students is instructional.

ADVANTAGE 6. Very limited. Advantage not exploited.

ADVANTAGE 7. Obtained.

CONCLUSIONS: While Advantages 1, 5, and 7 are exploited, those concerned with self-criticism (3), self-responsibility (2) and variety (6) are not. There is a clear predominance of Job Satisfaction weightings, which, without other weightings and consequent advantages could be interpreted as simply giving the learners a change of scene and activity with no clear pedagogical objectives.

SESSION 39

ADVANTAGE 1. Fully exploited. All weightings.

ADVANTAGE 2. Fully exploited. All weightings.

ADVANTAGE 3. Adequately exploited. 'A' weightings predominate.

ADVANTAGE 4. Fully exploited. Forty minute session. High interest and participation values, although single pace throughout on tape.

ADVANTAGE 5. Fully exploited. C.f. ADVANTAGE 2. Pedagogical communication encouraging self-criticism and self-responsibility.

ADVANTAGE 6. Not exploited.

ADVANTAGE 7. Adequate. Cassette hiss.

CONCLUSIONS: All Advantages except Number 6 (Variety) exploited. While all weightings are present, emphasis is on accuracy and self-responsibility (i.e. Job Satisfaction) in this session.

BEST COPY AVAILABLE

4 Final comments

I have argued that instead of just comparing the LL with other ways of teaching or learning languages, we need to evaluate how the LL itself is used. I have argued that we can see the rationale behind the LL in terms of three criteria, that it is faster, better, or gives more job satisfaction, and of seven advantages to be gained, potentially, in using the LL.

I would suggest that if the LL is not exploited in such a way as to fulfil the first two criteria, then the user must demonstrate how the LL is suited to the use being made of it in both technological terms and pedagogical terms. The models and grids which I have presented and described do have the potential to raise user-consciousness (especially as teacher and student checklists) through demonstrating clearly where improvements can be made. As I found in my studies, it is possible to use the LL fully and well in a very wide variety of ways - not just structural drills, listening exercises, or pronunciation work. In fact, I would say that its uses and usefulness are limited by only two factors: the skills and imagination of the teacher, and the degree of training, preparation and responsibility which the teacher can and is willing to give the learners.

Bibliography

Anderson, A. 1977. "How the language laboratory is used in schools". System 5, 1: 19-26.

Green, P.S.(ed). 1975. The Language Laboratory in School: Performance and Prediction. The York Study. Edinburgh: Oliver and Boyd.

Holec, H. 1971. "Laboratoire et efficacité". Mélanges Pedagogiques 1-17.

Lorge, S.W. 1964. "Language laboratory research studies in New York City High Schools: a discussion of the program and the findings". Modern Language Journal 48, 7: 409-419.

Savignon, S. 1972. Communicative Competence. Philadelphia: Center for Curriculum Development, Inc.

Smith, P.D. 1970. A comparison of the cognitive and audio-lingual approaches to foreign language instruction. Philadelphia: Center for Curriculum Development.

Vanderplank, R.N. 1981. "Using the language laboratory to develop the listening ability of adult learners of English by means of training in the perception of stress". Unpublished Ph.D. dissertation, University of Edinburgh.

LANGUAGE TESTING: MAKING A DECISION ON THE BASIS OF OBSERVED BEHAVIOUR AND 'TAPPED' KNOWLEDGE

Jan van Weeren (Cito, Netherlands)

0 Introductory remarks

In this paper I will present a rather general framework concerning the linguistic background of language testing. I will stress the fact that a test is not just a set of valid items, but that it is essentially an instrument with a specific function. Various aspects of tests will be discussed, such as the problem of gathering relevant information from testscores, in order to make decisions. A test for immigrant teachers is taken as an example. The results of this test will be related to the general framework presented at the beginning.

1 The nature of language proficiency

By saying that someone is proficient in a foreign language we imply that he or she is able to perform in that language, is able to show actual foreign language behaviour, in short, that he has a certain behavioural ability. There are several ways to look at this ability when it is necessary to measure language proficiency. It can be measured on the basis of actual authentic language behaviour, by determining the quality of specimens of authentic language behaviour. But it is also possible to consider this behavioural ability as a form of tacit, implicit knowledge which underlies authentic language behaviour, but which does not necessarily need to be evaluated through this behaviour. One can specify which tasks someone must be able to perform on the basis if his assumed tacit knowledge of a language, without these tasks necessarily taking the form of authentic language behaviour. In a third approach to this behavioural ability evaluation takes place by virtue of the individual's explicit knowledge about a language. A testee is expected to be able to state the rules concerning the use of the simple present and the progressive form in English, to rattle off the paradigma je suis, tu es, il est, nous sommes, vous êtes, ils sont in French, or to indicate with whom he or she is allowed to be on familiar terms in German. This explicit knowledge was traditionally considered as an important precondition for writing in a foreign language. In short: Evaluation of language proficiency as behavioural ability seems to be possible, firstly by observing authentic language behaviour, secondly by an examination of the tacit knowledge that underlies language behaviour and thirdly by testing the acquired explicit knowledge. In the following I will discuss these approaches successively.

1.1 Evaluation by means of authentic language behaviour

Evaluation of language proficiency as a behavioural ability by means of authentic language behaviour entails a method which is extremely face-valid. If you want to know if a person can take part in a foreign language conversation, you make him or her participate in it. If you want to know if he or she is able to read a text of a specific type, you put questions to him that test his comprehension of the text, or, alternatively, you make her write a précis. As a rule, there will be no objections to such an approach. It is clear and obvious that the procedure makes sense. Add to this that the method entails a positive backwash effect: as a result of the way of testing matters are trained in school that are actually required outside school.

However, when the testing expert comes in a number of problems will arise that increases the more the language behaviour under observation resembles the language behaviour that we meet in practice. Human beings usually do not read texts in order to answer a fixed set of reading comprehension questions, but in order to realize specific goals that can vary with the individual and the situation. Their reading comprehension can express itself in various ways. Distinct aspects of the text may be relevant for some, but not for others.

However, testing will require a certain amount of standardization. A conversation which is truly free and open is known for the sort of evaluational problems that are related to the reliability of rating. Regarding the free conversation there is yet another source of unreliability, that is: the testees themselves. Not everyone will be able to speak just as easily about any topic. One should consider factors as empathy and affective thresholds, factors that have very little to do with language proficiency (Underhill 1983).

Another problem is that of content validity: to what extent can observations of incidental language behaviour provide us reliable and complete information about the testee's behavioural ability, that the tester is to judge eventually?

1.2 Evaluation by means of tasks based on implicit knowledge

Specific tasks based on capabilities that are linked up with implicit knowledge of a language can be derived from traditional generative linguistics. This discipline operates on the basis of the concept of competence. Linguistic competence is generally defined as the individual's knowledge of the structure of his language. On the basis of this knowledge a native speaker is able to make judgements about his language system resulting from his 'true linguistic intuitions'. He is able to discriminate between well-formed and not well-formed sentences and to distinguish semantical differences and similarities in sentences that belong to his language.

Dell Hymes expanded the Chomskyan concept of linguistic competence to communicative competence. Competence remains underlying knowledge, but is expanded to include all the aspects of knowledge that affect communicative behaviour. On the basis of this knowledge a native speaker is able to judge not only the

grammaticality of sentences, but also the appropriateness of utterances. His knowledge does not confine itself to the language as a grammatical system, but is expanded to the use of the language. The native speaker knows whether and to what degree something is appropriate, that is, adequate, fortunate and successful in relation to a context in which it is used and evaluated.

Tasks that invoke capabilities linked up with implicit knowledge are very common in traditional language tests, not so much because of the fact that these tests are explicitly based on the theoretical implications of the concept of competence, as because of the necessary concessions one has to make on test-theoretical grounds if language behaviour is to be tested.
I will give you two examples of this type of task:

figure 1

HERE'S _____ IMPORTANT INFORMATION FOR ALL OUR CUSTOMERS

1) AN
2) ANY
3) EVERY
4) SOME

One might say that this item is testing the implicit knowledge of a selectional restriction rule (Modelltest VHS Zertifikat English, item 46).

figure 2

1) VOUS ACHETEZ SOUVENT CE JOURNAL ?
2) EST-CE QU'ELLE ACHÈTE LE JOURNAL ?
3) QUAND ACHETEZ-VOUS CE JOURNAL ?
4) VOUS ACHETEZ SOUVENT DES REVUES ?

All utterances are well-formed, but only one is appropriate in this context; the testee must possess some communicative competence (Standardprov French, Lindblad 1983, p. 57).

From a theoretical point of view the concept of communicative competence is too hybrid, although the concept of linguistic competence cannot be said to be theoretically uncomplicated, either. From a pragmatical point of view, however, the concept of competence gives support to many traditional language tests.

The main question is how test data obtained by means of items based on such tasks, relate to the behavioural ability that we aim at.

It can be hypothesized that the attained level of performence on these tasks gives us an indication of someone's language proficiency as behavioural ability. To put it in a different way: the obtained performance data will represent some measure of language proficiency.

A testing procedure which is based on tasks that a testee must be able to perform in virtue of his competence shares the problem of content validity with the evaluation by means of authentic language behaviour. Every testing procedure has to confine itself to a sample of possible performance. It is necessary to indicate or to test empirically to what extent such a sample is representative.

The hypothesis that testing based on tasks fitting the concept of competence on one hand and the evaluation of authentic language behaviour on the other hand will yield equivalent information about language proficiency as behavioural ability, can be tested empirically. This can be done by determining the concurrent validity of two alternative testforms. Clark (1972) found, for example, correlations from .82-.92 between the FSI-interview and a battery of objective tests for vocabulary and structures.

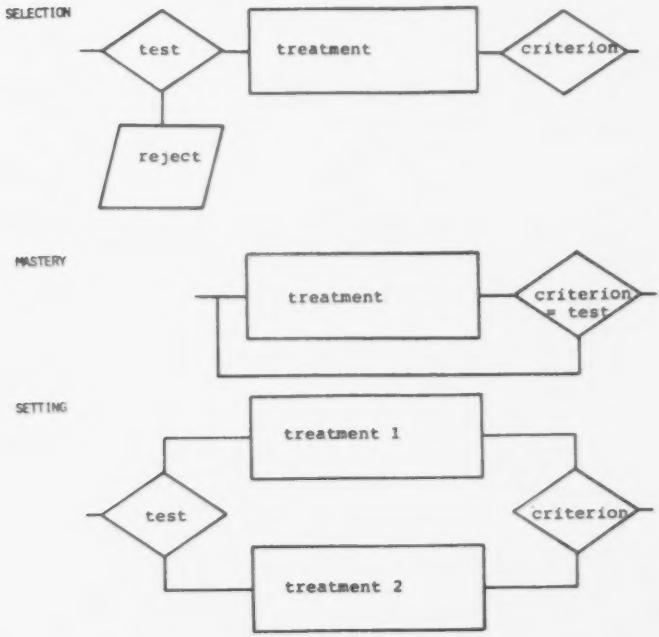
Tests consisting of tasks that the testee must be able to perform because of his implicit knowledge of a language, have the advantage of the possibility of an objective form. An exclusive use of this type of test in instructional settings, however, will inevitably carry with it the disadvantage of an undesirable backwash effect, unless particular measures concerning the curriculum have been provided against this effect. Otherwise educational activities will focus on learning tasks of an abstract nature and training of the actual use of the language will be neglected.

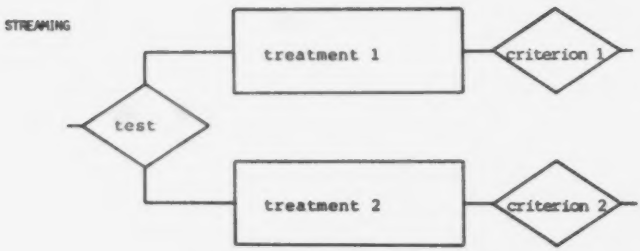
1.3 Evaluation by testing explicit knowledge

In respect to the form of testing that focuses on explicit language knowledge we can be brief. As early as half a century ago it was stated that knowing about a language and knowing of a language are two different things. Explicit knowledge of rules is neither a necessary, nor a sufficient condition for a successful and all-round language use. The first is proven by native speakers of a language who did not receive any formal schooling in linguistics, the second by grammar school pupils who, though having a thorough command of explicit rules, cannot speak a word of Latin or Greek. A teaching process that prepares for this kind of testing should be considered as philology, rather than as language teaching.

2 Test functions

It is remarkable that among applied linguists the form and theoretical background of test items are the main topics under discussion. They mistakenly call a set of test items a test, whereas a test is essentially a different thing, namely a measuring instrument that provides useful information to be used in decision making. In this respect linguists are a bit like instrument makers who talk about a clinical thermometer and think that the thermometer is good if the column of mercury is in different positions with different persons and in the same position with the same person under conditions of repeated measurement. However, a clinical thermometer is only fit if it provides information that enables the user to decide if he or she should stay in bed, should consult a doctor, should take medicines or has reached a period of fertility.
In the same way a test must provide information that enables us to decisions of the following kind:





The concrete test I would like to discuss meant to support a process of making decisions of the first kind. The testees were groups of teachers who had come from various countries: Turkey, Morocco, Spain, Italy, Jugoslavia, Portugal, Greece and Tunesia, to teach their native language and culture to groups of children with the same cultural background in primary schools. The treatment offered was a special course that would make them fully qualified teachers in primary schools. It was observed that these teachers were not integrated into the teaching staffs, mainly because of the fact that they were not qualified to teach any other subject but their native language and culture. Neither were they allowed to teach other pupils than those of their own culture. In order to further the integration and to widen the scope of these teachers the course would give them the oppurtunity to acquire a full qualification. The criterion in the flowchart would consequently cover the objectives of a regular college of education.

To set an entrance level it was required that the candidate members of the course had a satisfactory oral command of the Dutch language so that they could communicate rather fluently.

On the authority of the inspectorate this requirement was tightened up. To a certain extent command of the Dutch language was required at the level of pupils in the last form of higher level secondary schools. The inspectorate had its reasons for this: the course would only take two years and after finishing it these teachers would be formally qualified to teach the Dutch language to Dutch children in primary schools!

On the basis of these requirements global selectional criteria were set.

Cito, that is the Dutch National Institute of Educational Measurement, was charged with their operationalization.

3 Choice and development of the instrument

With the theoretical considerations put forward at the beginning of this paper in mind it was tried to obtain information about the language proficiency of the testees in two different ways: firstly by eliciting and evaluating authentic language behaviour and secondly by administrating items that would measure the underlying implicit knowledge.

3.1 The oral proficiency subtest

For the evaluation of language behaviour the following type of oral proficiency test was designed. It consists of about 15 stimuli requiring the testees to play a certain part in a situational dialogue. The stimuli are presented verbally. In most of the cases a response is prompted by means of a picture, for example:

AT THE POLICE STATION

One day you discover your wallet is missing. You go to the police station.
Good morning, an officer says. What can I do for you? /.../ Could you tell me what was in your wallet? (picture 1) /.../ Have you any idea where you could have lost your wallet? (picture 2) /.../ etcetera.

These stimuli are followed by circa 15 general questions about their life in Holland.

The responses were each judged on intelligibility/appropriateness and correctness. To each response a maximum of 6 points could be assigned. The maximum meant that a response was perfect: appropriate in the context given, perfectly intelligible and completely correct in a grammatical sense. 5 points were assigned if the response was intelligible and appropriate, although some minor mistakes were made. If it took some effort to understand the response as a result of certain important lexical or grammatical mistakes, 3 points were assigned. If a response was either not forthcoming or not intelligible or not appropriate at all, 1 point was assigned. In all cases one point was deducted if repetition of the stimulus was necessary.
In this rating system a sum total of 180 points with 30 items means that all responses are perfect: intelligible, appropriate and completely correct. There is a rapid decrease in degree of correctness: a score of 150 points means that the average response is appropriate and intelligible, but no more than that. With lower scores a general appropriateness and intelligibility is preserved for a relative long period. Understanding takes some effort with scores below 140. With a score of 120 or 110 points several misunderstandings will arise and communication will start to break down.

Testing procedures like this one based on authentic language behaviour, have the additional advantage of enabling us to apply natural, that is to say, common sense criteria when cut-off-scores have to be determined. Criteria such as 'oral command of Dutch on a near-native level' or 'being able to make oneself understood' can almost immediately be translated into a score on the oral proficiency test. An example can be taken from the rating system of the FSI-interview: a naturel criterion related to overt language behaviour corresponds directly to a specific test score.

In the case of the teacher and the test mentioned above the requirement of oral proficiency was operationalized as a sufficient degree of appropriateness and intelligibility in oral expression as measured by the test. With some safety margin the cut-off-score was set on 145 points.

3.2 Testing implicit knowledge

For the measurement of implicit language knowledge three subtests were composed. One of them aimed at vocabulary and structures: at the implicit knowledge of rules concerning the derivation of words, synonymy, the use of conjunctions and the building of sentences. It consisted of multiple-choice items of a rather straightforward type. The multiple-choice items in the second subtest focused on the orthography of changeable and unchangeable words. This part leaned heavely on the quality of literacy training in Dutch. Therefore it was more like an achievement measure than an indicator of language proficiency. The third subtest consisted of a cloze test in multiple-choice-format. In fragments taken from transcriptions of newscasts every 5th or 7th word was deleted.

Distractors were obtained by administrating an open version of the test with foreign university students. With regard to the underlying knowledge measured by this cloze test it can be concluded that widely different aspects were involved. Apart from lexical fluency the test required from the testees that they could process divergent structural and logico-semantical information in the context.

3.3 Setting a cut-off-score for tests based on implicit knowledge

There are no natural, that is, common sense criteria for the determination of a cut-off-score in tests that are focused on linguistic and communicative competence. It is not possible to formulate such a criterion in terms of observable and functional language behaviour testees are expected to show. For lack of natural criteria the setting of a cut-off-score will depend on the test itself. In consequence, the outcome is essentially arbitrary. I will sketch two current standard procedures. According to the first method a group of testees sets its own norm. After the administration of a test the mean score of the group is calculated and the cut-off-score is set somewhere below the mean (for example at a distance of twice the standard error of measurement). In the second method a group of experts is to determine what they consider as a passable or not passable perfomance on the test. This can be done in an informal way, by discussion and consensus, or following to some formal procedure (for example that of Nedelsky, Angoff or De Groot). A combination of both methods is applied in centralized final examinations in Dutch secondary schools. A preliminary cut-off-score is set on a intuitive basis. Depending on concrete testresults an adjustment is made if necessary. The percentages of pupils that pass or fail should approximalety equal those of previous years. This is a proper method if testpopulations are sufficiently stable each year with regard to their educational background.

3.4 The choice of a reference population

In some cases there may be an opportunity to set a norm in an empirical way by introducing a well-defined reference population. To give an example, if the language proficiency of a target population should equal the proficiency level of the average

succesful foreign university students, this norm can be obtained by pretesting.
This solution presented itself with the test for the foreign teachers. I remind you of the fact that not only a functional oral command of the second language was required, but also some practical knowledge of Dutch on the level of pupils in the last form of higher level general secondary schools. These are five-form schools following primary education for pupils between the ages of 12 and 17. These provided us with a clear-cut reference population.

The intended norm was defined as the score that anyone could reach who was good enough to reach the last form. At first sight finding this norm seemed very easy. However, we were confronted with a slight complication.

Pretesting would depend on voluntary participation. The test would not have any consequences for the pretest population. Thus there was a change that some pupils would take it less seriously. Apart from sabotage we had to consider the fact that one pupil or the other would have an off-day. For that reason it was rather dangerous to define the norm as the lowest score of the selected pretest population. Therefore it seemed reasonable to take that score as a norm that was two standard deviations below the mean score. With the expected distribution of scores this norm would be reached by circa 95 percent of the pupils.
A pretest was carried out with the following results:

RESULTS OF THE REFERENCE POPULATION (1983)

	n	KR20	$\bar{p}$	%< M-2σ
vocabulary and structures	111	.50	.94	2,7
orthography	111	.85	.89	6,3
cloze	249	.92	.91	4,4

The high reliability coefficients for orthography and cloze gave food for thought. It might very well have been that some pupils did not do their utmost. The item-analysis of the cloze test revealed that several items were skipped at the end, that is, no answers were given. If this occurs systematically, that is, if the same testees skipped each of these items, reliability is flattered. Pretest results of a similar test in 1984 are more realistic from this point of view:

RESULTS OF THE REFERENCE POPULATION (1984)

	n	KR20	$\bar{p}$	%< M-2σ
vocabulary and structures	193	.45	.96	0
orthography	193			
cloze	193	.63	.98	0

Still, reliability of the cloze test is quite high, in view of the impressive percentage of correct answers. However, extremely low scores did not occur anymore.
After the pretesting of the reference population norms were set following the procedure described.

4 Testresults

Administration of the test with 87 immigrant teachers yielded the following results:

	number of items	KR20/α	p̄
vocabulary and structures	10	.75	.55
orthography	20	.86	.56
cloze	100	.96	.52
speaking prof.	30	.96	.67

Obviously the speaking proficiency test was much easier than the other subtests. No wonder, in view of the fact that only a Basic Interpersonal Communication Skill (Cummins 1979) was involved, without the requirement of native speaker proficiency.
The rating of the elicited responses on the speaking proficiency test was carried out by rotating pairs of raters. Each rater assigned his scores independently. The inter-rater-reliability was surprisingly high. To a certain degree this could be explained by the heterogenity of the testpopulation.

pairs of raters	R
A/B	.98
A/C	.98
B/D	.97
B/C	.96

The following correlation coefficients between the various subtests were found:

cloze x vocab. & struct.	.78
cloze x orthography	.74
cloze x speaking prof.	.71
vocab. & struct. x orthography	.70
speaking prof. x vocab. & struct.	.68
speaking prof. x orthography	.65

5 Conclusions and discussion

On the basis of these modest results we could not reject the hypothesis that the measurement of language proficiency can be done by evaluation of language behaviour as well as by the testing of implicit knowledge. Instruments of both types do provide equivalent results. How these instruments relate to each other may appear from the following survey:
9 candidates with a sufficient score on the speaking proficiency test passed the vocabulary and structures subtest as well as the cloze test. Only one of the candidates scoring just below the cut-off-score of the speaking proficiency test (scores ranging from 140 to 144 points) passed both the vocabulary ans structures subtest and the cloze test. Another candidate passed the cloze test only. Below 140 points just one candidate passed both subtests. If we had based our selection decisions exclusively on the subtests vocabulary and structures and the cloze test, only three misclassifications would have resulted in that three candidates out of 87 with an insufficient speaking proficiency would have passed the test.

With regard to the possibility of misclassifications the results of the subtest orthography were more problematic. Of the candidates that passed the speaking proficiency test 16 passed the orthography test as well, but among the candidates just below the cut-off-score there still were four of them that passed the orthography test, and with lower scores there even were six!

It is obvious that orthography is linked up with language knowledge, but to a large extent it can be acquired as an isolated system, especially if it is confined to general rules. Those who passed the orthography subtest with a deficient speaking proficiency, might have acquired their knowledge by their unflagging energy in language classes.

This illustrates the necessity to take the factor language background into consideration when testing language proficiency (Cziko 1984). Language background refers to the type of contact the testees have had with te second language and the opportunity they have had for acquiring the various aspects of the language. If this language background consists of a language course where writing skills are emphasized, language proficiency might be flattered if these skills are much represented in the evaluation procedure. If, on the other hand, the various aspects of a language are trained in a more balanced way, distinct subtests will represent language proficiency as a whole more adequately.

Bibliography

Hymes, D. (1972), 'On Communicative Competence', in J.B.Pride & J.Holmes (eds.), Sociolinguistics, Harmondsworth, 269-93.

Underhill, N. (1983), 'Commonsense in Oral Testing: Reliability, Validity and Affective Factors', in E.W.Stevick a.o. (eds.), On TESOL '82: Pacific Perspectives on Language Learning and Teaching II, Tesol Washington, 125-39.

Backlund, P. & J.Wiemann (1978), 'Current Theory and Research in Speech Communicative Competencies: Issues and Methods', paper presented at the annual convention of the American Educational Research Association, Toronto.

Cziko, G. (1984), 'Some Problems with Empirically-based Models of Communicative Competence', Applied Linguistics, Vol. 5, no.1, 23-38.

Lowe, P. (1976), The Oral Language Proficiency Test, Interagency Language Roundtable, Washington D.C.

Deutscher Volkshochschul-Verband, Pädagogische Arbeitsstelle (1980), Das VHS-Zertifikat Englisch, Bonn.

Lindblad, T. (1983), 'The Swedish System of Final Examinations in Modern Foreign Languages', in J.van Weeren (ed), Practice and Problems in Language Testing 5, Cito Arnhem, 51-63.

Cummins, J. (1979), 'Linguistic Interdependence and the Educational Development of Bilingual Children', Review of Educational Research 49, 222-51.

CULTURAL KNOWLEDGE AS PART OF LANGUAGE SKILLS IN CROSS-CULTURAL COMMUNICATION

Pertti Widén (Turku School of Economics and Business Administration)

Communication like every human activity is constrained by individual and cultural factors. In fact we cannot know how the receiver has understood the sender's message in a communication process, but the idea of communication must involve a common understanding. Bearing in mind the fact that communication is not merely language (in a strict sense), we can better uncover the cultural constraints of communication ("Culture" is used here in a very wide sense as "the ways of a people").

Communication is a two-way-process and only successful when the receiver has understood what the sender intended. There is an intention, a message, which has to be transmitted to the receiver by the sender in a code system that is common to both actors in communication in order to influence the receiver in a particular way. Roughly speaking there are three kinds of factors to be considered in a communication process:

1. characteristics, habits, actual psycho-physical determinants, expectations etc. of the sender;

2. characteristics, habits, actual psycho-physical determinants, expectations etc. of the receiver; and

3. situational factors (context) (which may also be different from the respective viewpoints of the sender and the receiver).

Here I would like to underline the synergy and the interdependence of all factors in a communication process and also the importance of the receiver in the communication process. The receiver is not a block without reactions (even if in some cases the receiver may appear difficult, or unwilling to communicate or to be influenced; no reaction is also a reaction and will have an influence on the behaviour of the sender). In fact, communication is a dynamic process, not static, but a complex situation where everything matters.

Each communication partner has their own interaction competence (Oksaar 1984: 30-) in performing and interpreting verbal and non-verbal activities in an interaction situation. Initially familiar models of communication strategies are applied if there is no better alternative. We can therefore say that each partner in communication has their own combination of cultural background and individual factors when they enter a communication situation. Even perception and interpretation of the communication situation is a cultural matter and may thus differ significantly from the particular viewpoints of the communication partners. Awareness of the cultural constraints in communication is crucial in cross-cultural understanding. To

improve communication, one should therefore consider the cultural ties involved in communication and utilize this knowledge as a partner in communication. In fact this consideration involves two things: firstly, an understanding of the behaviour and habits of the communication partner including an understanding of the communication partner's way of perceiving and interpreting the communication context; secondly, it is necessary to be ready to explain one's own behaviour, and habits and one's own way of perceiving and interpreting the communication context. Very often the latter point is forgotten. If we explain (whenever possible) why we behave and react in a particular way, explanations might clarify a situation and perhaps also make the matter (and ourselves) more interesting to the communication partner. In this connection, the amount of the information that has to be explicitly coded will vary from culture to culture. As we know from the HC/LC scale of Edward T. Hall (Hall 1977: 91-), a high-context (HC) communication or message is one in which most of information is either in the physical context or internalized in the person, whilst very little is included in the coded, explicit, transmitted part of the message; a low-context (LC) communication is just the opposite, i.e. the mass of the information is vested in the explicit code. The different scale positions of the communication partners may lead to serious misunderstandings in cross-cultural communication if the partners are not aware of these differences, as, for example Stella Ting-Toomey (Ting-Toomey 1984) has shown for American ("low-context-culture") and Chinese/Japanese culture ("high-context-culture") at the World Congress of Applied Linguistics in Brussels in August 1984.

Let us now consider some of the characteristics of Finnish-German communication as examples of factors which we can study in order to find out something about the communication barriers in cross-cultural communication.

One of common parameters that varies from culture to culture is the time factor, the attitude to time. In my investigations (Widén 1983) I could distinguish similar differences between Finns and Germans as those found by Stella Ting-Toomey between American (LC) and Chinese/Japanese (HC) culture. There are gradual differences and different accentuations, but nevertheless these are differences that influence the whole communication process.

First, life seems to be more hectic in Germany than in Finland, time has become a kind of currency, "time is money". This attitude makes it possible for example for a person to be interrupted in a TV discussion if one second more time is needed for finishing a sentence than is provided for; time is more important than politeness (in fact under those circumstances the interruption is not regarded as impolite). In Finland this would only be allowed in a heated TV debate. This is not the best example because other factors may also play a role and influence the communication process and the communication behaviour, for example role, status, dominant behaviour in the interaction etc. (cf. Hämäläinen 1982, 141-144).

A second example is time or the length of the reaction in a communication process. Here - generally speaking (neglecting regional and individual differences) - Germans seem to be quicker in response than Finns. "It takes too much time to get an answer from a Finn", is a comment that is very often heard from a German. In this instance it is not only the attitude to time that matters but also a difference of opinion about the answer. A German would like to hear immediately if the communication partner has understood the message, a Finn however would like to express a final response only, and he does not see the value of an immediate intermediate reaction as a confirmation, unless made aware of this as an aid to the communication process. A typical German reaction in a situation like this (with no immediate answer from the Finnish partner) would be for the German to make various suggestions, to "help" the communication partner. To the Finnish partner this could complicate the situation, the help may be misunderstood and the suggestions thought of as arrogance from the German side. And if mutual empathy does not suffice, the communication may suffer because of this interference. Empathy is also needed for other aspects of the situation; for example the slow reaction from a Finn must not make the German partner believe that the Finnish partner is dull just because of a slow reaction.

Attitude to time also influences the organization of the discourse. A German (and also an American) will tend to verbalize the whole process, including thinking, and will use at least reactive (if not active) comments when "listening aloud". A Finn on the other hand (and also a Japanese) will tend to verbalize the product of thinking only, assuming that the communication partner understands without saying that the silence is used for thinking. Such a discourse structure would more like a monologue. This is also typical of Turkish discourse habits, and this has caused difficulties between Turkish people and the Germans in Berlin e.g. (cf. Kuure, Kärkkäinen, Sipola, and Piirainen, 1984: 17; for American/Japanese comparison see Okabe 1983). We can see therefore that different attitudes to time can influence the discourse structure so that the whole communication process may suffer from this in cross-cultural communication.

A common attitude in Finland is: "A Finnish man neither speaks nor kisses". This means that a Finn is not used to verbalize or to show feelings, but on the other hand it also shows the attitude of a Finn towards verbalized communication: obvious things are not worth mentioning. The problem in this connection is naturally: What is obvious? A discrepancy between the particular expectations of the sender on the one hand and of the receiver on the other hand may here cause problems, not only interculturally but also intraculturally. Anyway, the old cultural habit that only important matters, real messages, are worth mentioning, are worth articulating, naturally also influences the form of Finnish communication: a typical Finnish discourse consists of statements only, questions occur relatively seldom. Furthermore, "small talk" is of little importance. So it is no wonder that it is sometimes difficult to maintain

conversation in Finland. Other languages and other cultures may have the custom not to be quiet in company. This is an important matter to be noticed when teaching the language concerned in Finland. Without content - i.e. having something to say, a definite speaker intention - there can be no structure.

It is not only "speaking" that counts, but also "kissing", i.e. the various forms of non-verbal communication, which are also important in communication. The French can hardly talk without using gestures, a Finn uses body language very little. And this aspect is also a task for language teaching, if we want to teach communication in a particular language, not just words and grammar. In many cases these things are interrelated: the expressive structure of one language may need more contextual support than that of another language (cf. HC/LC-discussion above). The lack of "eye contact" of the Finns could be related to the fact that the Finnish language does not use paralinguistic means and prosodic features (intonation etc.) to any extent, so it is not necessary to follow mouth movements or the countenance of the communication partner. Under these circumstances it is possible that not looking at the partner may be regarded as polite behaviour. This is the case in Finland. Many foreigners wonder how Finns can manage not to look at anybody in lift that is chockfull; they do not understand that this is a kind of Finnish politeness. To the Finns, it would be an insolence to look at a stranger so obviously that he would notice it. It is not only the Finns who react in this way. For the English privacy is an important matter which is widely respected by other people ("An Englishman's home is his castle"); for the Japanese, whose everyday life including all chores is very public, feelings are private and they do not want to show them publicly.

But not only should the degree of non-verbal communication and its relative significance in comparison with verbal communication be considered, but also the importance of consistency between verbal and non-verbal communication. Otherwise there may be confusions in communication because of inconsistency, the receiver will not know what to believe (according to various investigations in such cases the receiver tends to believe the non-verbal rather than the verbal message).

Another example of the interrelation of language and culture is the use of names. In English mentioning of names has a very important role because of the existence of just one address pronoun "you", although in practice this is also a useful elliptical form when elements of uncertainty exist. In various cultures mentioning of the name of the communication partner is a demand of politeness. We do not say "Guten Tag" in German, but "Guten Tag, Frau Müller", we do not say "Bon jour" in French, but "Bon jour, madame/Bon jour, monsieur". In Finland mentioning of names in this fashion is not so common: Finns are perhaps more influenced by the past, more archaic. Among many primitive people, the mentioning of a person's name is an indication of some sort of power over the individual. This kind of false modesty in terms of not mentioning one's name

or not asking the communication partner's name (which is not only a habit of the Finns) can naturally lead to misunderstandings or confusions in a communication situation. Only if the communication partner is used to asking the partner's name (as for example Germans), is there no harm to further communication. In many cases, communication problems can be avoided by cooperation of communication partners in a communication situation.

One of the most important things in cross-cultural communication is adequate presentation, argument, stating facts and reasons so that the communication partner can understand the message. In this connection we have to separate between content and structure, both of which vary from culture to culture. The goal of right presentation (argument) is to find some kind of "common language" for both communication partners. The "common language" is a very difficult area. Els Oksaar (Oksaar 1976) speaks of the social dimension of linguistic variation and of the linguistic dimension of social variation in this connection when describing interpersonal communication. The easiest aspect here - and even this is complex - are the connotations of words - those secondary meanings of words which in communication tend to become primary meanings. For example, the word "work" has a different meaning for those who have a job compared to those without a job. Also the attitude to work varies considerably from culture to culture. In an international comparison, the European Values Study Group found out recently that Finns, like other Protestant and partly Protestant nations, still regard work and the satisfaction derived from work as a central base for life and feelings - quite opposite to Catholic nations or countries like South Korea or Japan which were also represented in this investigation (cf. Lotti 1983). Current values in a culture can also influence the actual meanings of words and expressions. And so too can different life experiences and customs. For a German "sport" means mainly football, for a Finn (in summer) mainly athletics (excluding football fans, of course!).

An example of a larger argumentative content is the attitude to money. Expensiveness, or lack of money, are widely accepted refusing arguments in several cultures, in other cultures it is not polite to refer to anything which has to do with money. In Germany, a city guide will always remember to mention the cost of construction of an important building, something that always surprises a Finn. And the first question of the German visiting Turku Castle (which was destroyed during the war and then restored) is usually, what was the cost of the restoration. A Finn would never ask this sort of question because it would not be polite.

It is a part of the task of a language teacher to point out these things. It is important to know the life circumstances and the cultural background of a communication partner in order to be able to use facts and expressions that can be understood in the right way. Finns often know quite a lot about other countries and nations, and in many cases they assume (quite wrongly) that foreigners know as much about Finland. Usually foreigners do not know

anything about Finland; Finland was never an important school subject, Finland has not played a large role in world history, and Finnish is not a worldwide language. A Finn must therefore give a great deal of basic facts about Finland in order to offer an adequate picture of Finnish circumstances.

Not only are the contents, the argumentative nature and kind of presentation, important, but so too is the form and order of presentation, the particular discourse pattern. We have discussed above the nature of discourse patterns and how they vary from culture to culture. International comparisons of student essays have shown how they also vary from culture to culture. Such essays have shown for example that Finnish school children can tell stories, describe things and events, but have difficulties with justifying claims, stating reasons and giving grounds for their opinions. This result seems to be connected with Finnish discourse characteristics which tend to use laconic statements only. At the AILA 84 World Congress, it was also interesting to hear that the best essays in a comparison of Finnish, British, American and German student essays seemed to follow a similar discussion order, whilst the essays with the lowest scores showed considerable cultural differences; in this group the Finnish and German essays seemed to form a group of their own, separate from the other essays (cf. Connor 1984). Under these circumstances, unless the composition of an essay in a foreign language is a special subject of foreign language teaching, I would not recommend a free essay form for language testing purposes. This is mainly because there are so many aspects to be considered, not just language structure (grammar) and vocabulary. It is a cultural matter whether one has something to say or not (see above: content) and if an individual has the creative ability to be able to put it into words (not everyone can do this even in the mother tongue!). And this cannot be done in a foreign language using familiar mother tongue strategies. It is not only pupils who have difficulties with the free essay form, but also teachers and those evaluating essays. Even the sense of humour varies, not only from culture to culture, but also in the various sections of a language community. If the command of the written form of a language has to be tested, there are many other more adequate forms of test procedures (guided writing, summaries etc.) which can be used instead.

As we have seen, everything that is said in a communication situation matters. This is something that we also know from the classic formula of Lasswell: "Who says to whom, what, when, why, and with what effect". But there are also things that are not said that count. There are topics which cannot be discussed in certain cultures. For example a Finn prefers not to say anything about political or religious beliefs - these are matters which are considered to belong to everyone's private life in Finland. The opposite is true to a German, who, even unasked, likes to announce political opinions. This obvious need to discuss all possible things is not shared by a Finn. I do not know if the Fulbright professor who came to Finland in order to investigate the silent breaks in the talk of the Häme-Finns (who are famous for their taciturn nature and chariness

with words) has completed his work but the results might be very interesting from the cross-cultural point of view. This sort of silence should be put into a larger communicative frameword, e.g. in terms of HC/LC-culture. I have noticed that several Finns are also unhappy about too long silent passages in company as well but cannot find anything to say because of inadequate communicative strategies. In some cases they use surrogate activities instead of talking such as sighing, smoking or (in an intimate company) yawning: so it is important to understand that a yawning Finn is not necessarily a bored Finn.

Above we have also mentioned several interaction determinants which can influence the communication process, things such as distance and dominance, status and role. People can have several roles in a communication situation. In cross-cultural communication there is always at least one person with the role of a foreigner. The fact of being the foreigner is not necessarily a barrier to communication (provided that there is sufficient command of a common language), the role of the foreigner can also be an advantage: in many cases a foreigner is not supposed to have a full command of language or culture. In Japan (Göhring 1980) a foreigner is not expected to have Japanese manners, whereas in Germany a foreigner with full command of German would be admired.

An adequate approach to cross-cultural communication always requires empathy and mutual cooperation. No one can master cross-cultural communication perfectly. Still, I am convinced that efforts in this direction will always be rewarded. It is not only a question of a particular communication situation, a happy end to business negotiations etc., but also of international understanding as a whole. Cultural knowledge as a part of language skills in cross-cultural communication is an important contribution to a better world-wide understanding.

BIBLIOGRAPHY

For many clarifying insights and hints I am grateful to Mr. Timothy Rawlins, who - being a practising expert in cross-cultural communication - has also reflected on these questions theoretically.

Connor, Ulla. 1984. "Argumentative patterns in student essays: a cross-cultural study". In Jan den Haes and Jos Nivette (ed). AILA Brussels 84 Proceedings, vol. 2. 541-542.

Göhring, Heinz. 1980. "Deutsch als Fremdsprache und interkulturelle Kommunikation". In Alois Wierlacher (ed). Fremdsprache Deutsch. UTB 912. Fink München. 70-80.

Hall, Edward T. 1977. Beyond culture. Anchor Books. Doubleday etc.: Anchor Press.

Hämäläinen, Simo O. 1982. "Mikrotason sosiolingvistiikka -keskustelun kielioppia". In M. K. Suojanen and Päivikki Suojanen (ed). Sosiolingvistiikan näkymiä. Helsinki: Gaudeamus Kyriiri Oy. 132-166.

Kuure, Leena, Kärkkäinen, Elise, Sipola, Seija and Piirainen, Arja. 1984. "Tekstin ja diskurssin tutkimus palvelee kielenopetusta". Tempus 7: 15-18.

Lotti, Leila. 1983. Suomalaisen ystävä. Helsinki: Markkinointi-Instituutti.

Okabe, Roichi. 1983. "Cultural assumptions of east and west: Japan and the United States". In William B. Gudykunst (ed). Intercultural communication theory. Current perspectives. International and intercultural communication annual, vol. 7. Beverly Hills etc.: Sage Publications. 21-44.

Oksaar, Els. 1976. Berufsbezeichnungen im heutigen Deutsch. Soziosemantische Untersuchungen. Mit deutschen und schwedischen experimentellen Kontrastierungen. (Sprache und Gegenwart 25). Düsseldorf: Schwann.

- 1984. "Fachsprachen, internationale Kompetenz und Kulturkontakt". In H. P. Kelz (ed). Fachsprache. Bonn: Ferd. Dümmlers Verlag. 30-45.

Ting-Toomey, Stella. 1984. "Intercultural misunderstanding: an interpretative analysis". In Jan den Haese & Jos Nivette (ed). AILA Brussels 84 Proceedings, vol. 1. 259-260.

Widén, Pertti. 1970. "Das Präteritum in der deutschen Gegenwartssprache. Imperfekt und Perfekt aus der Sicht der Sprachkommunikation". Linguistik und Didaktik 4: 283-287.

- 1971. "Kuuntelu puuttuu kielistudiosta". Helsingin Sanomat 11.7.1971.

- 1972a. Kielistudio-ohjelmat ja kielikommunikaatio. Turku: Turun kauppakorkeakoulun julkaisuja A1-4: 1972.

- 1972b. "Sprachlabor im Fortgeschrittenunterricht". Neue Unterrichtspraxis 1: 28-29.

- 1977a. "Kehnot käännökset väärää säästämistä". Kauppalehti 14.7.1977.

- 1977b. "Mikä on virkatodistus? Eli kansallisten termien kääntämisestä". Kääntäjä-Översättaren No. 7.3.

- 1982. "Erityisalojen tekstin ymmärtämisestä. Kokemuksia ja kommentteja Eindhovenin LSP-symposiumista". Kielikeskusuutisia/ Language Centre News No. 6: 15-19.

Widén, Pertti. 1983. "Kommunikatiivisuus on valttia Saksan markkinoilla". Mark 10: 46-47.

- 1984a. Interkulturelle Verständigung am Beispiel der finnisch-deutschen Kommunikation. (Duisburg. Forthcoming.)

- 1984b. Kulttuuri osana kielitaitoa yritysten kansainvälistymisessä ja sen huomioonotto kauppatieteellisen alan kielten-opetuksessa. (Jyväskylä. Forthcoming.)

CLASSROOM ENGLISH FOR GERMAN SCHOOLS

Bernd Voss (University of Bielefeld)

1 Introduction

The present report is an attempt to present some of the results obtained in the course of an inquiry into the language which native speakers of English would use to refer to selected aspects of the German classroom situation. It is part of a larger project concerned with the development of teaching materials for the field of "classroom language", for German teachers of English (1).

By "classroom language" we mean the language used by teachers and pupils to conduct classroom business. Our problem is that not all aspects of the German classroom situation have direct equivalents in the English-speaking world: some are rather specific and characteristic of schools in Germany rather than elsewhere. If one maintains, as we do, that using the target language for as many aspects of a foreign language lesson as possible can make an important contribution to closing the credibility gap between classroom and reality, ways need to be found to also talk about the characteristically German aspects of the classroom situation in English, in an acceptable way.

To avoid jumping to conclusions, or to be dependant on individual opinions only, 3o native speakers from various parts of the English-speaking world were asked to paraphrase in English a number of situations as they might arise in German classrooms.

The present paper is a report on this inquiry. After a brief sketch of the background and a description of the method used we shall present and discuss some of the results obtained in the course of this investigation before, finally, suggesting how this information can be used for the proposed teaching materials.

2 Background

Recent analyses of the role of language in classroom communication (cf. survey in Voss 1984) have shown that classroom discourse is not only teacher talk, but includes pupil-teacher and pupil-pupil interaction as well. More importantly for our present purpose, the realisation has grown that a considerable proportion of the verbal interaction in the classroom, if we consider a lesson as a whole, is focussed on fields other than the explicit lesson topic itself: A child might have his birthday, the next test is to be announced or postponed, there are not enough worksheets to go round and two children must be asked to share a copy, a pupil cannot find his homework notebook, has not done his homework, needs to be told off for misbehaving, cannot see the blackboard properly, complains about a neigh-

bour disturbing him, the whole class needs to be organised in groups for the next stage in the lesson, there is no chalk left and a child has to be sent to the janitor or secretary to get some more, etc.. All of these are aspects of a lesson involving language in a real communicative sense.

An attempt to systematise the various fields of reference of the verbal interaction occurring in the context of a lesson shows at least the following six fields of classroom discourse, (i) the topic itself, which obviously usually dominates the verbal interaction in the classroom, (ii) the real life situation of the teacher and the children in the classroom, (iii) the language, perhaps in the context of an explanation or a correction, (iv) discourse, in that teacher and pupils negotiate understanding and clear up misunderstandings, (v) organisation covering any of the many organisatory moves needed to keep classroom business going, and (vi) discipline (cf. fuller survey in Voss 1984).

Foreign language teachers do not necessarily accept all of these fields of classroom discourse as areas to be covered in the foreign language. While the first - and possibly also the third - tend to be somewhat automatically included in the foreign language parts of their lessons, many of the others are often not considered part of the lesson "proper", and are therefore dealt with in the mother tongue.

However, such a distinction between what is suitable for foreign language discourse and what is not has rather serious implications. It is easy to see that the fields often excluded in this way from being covered in the foreign language are in fact those which reflect aspects of the actual situation in which the teaching is carried out, i.e. the classroom situation itself. However, the classroom situation is the only genuine situation that institutionalised language teaching commonly provides. There is little doubt that the more these areas are excluded from foreign language discourse the more the learners must inevitably be left with the impression that the foreign language is only suitable for the verbalisation of the necessarily fictitious worlds of textbooks and other teaching materials. We would therefore argue very strongly in favour of an inclusion, as far as conceivably possible, of the actual classroom situation in all its aspects in what is dealt with in the foreign language. This would not only take advantage of an immediately relevant practice field in which language is used for real communicative purposes, but it would also help to close the credibility gap between classroom work and real life which otherwise would be inevitable.

While it may be easy to agree on the desirability of this approach in principle, the problems of putting this concept into practice are considerable. Teachers tend to receive little preparation for the classroom-related aspect of their foreign language command, within the framework of their training, nor do there seem to be many useful materials available to teachers interested in educating themselves in this respect on their own (cf. however, Hughes 1981, as a notable exception). In addition, however, what may even be more fundamental as a problem is the realisation that many aspects of the classroom situation of a particular country are in fact rather difficult to express in a foreign language at all.

There are two major reasons for this difficulty. The first is connected with the different ways in which languages tend to conceptualise their surroundings: what may be easy to say in one language may have to be paraphrased in the other, because a directly corresponding term (and possibly also the concept that goes along with it) may not be available. A German child who cannot see properly what it says on the board because the light is reflected on the shiny surface so as to blur the chalk writing will want to say Die Tafel blendet, but will find no easy way of saying this in English (nor will the teacher be able to suggest one) because blenden in this sense cannot be directly translated into English, and a paraphrase such as Can I sit elsewhere? I cannot see the board properly may have to be one of various situationally appropriate solutions to this problem.

The second major reason for the difficulties often experienced with the verbalisation of the various aspects of the classroom situation in the target language has to do with the fact that foreign languages are typically taught in the school setting of the home country of the learners and not in the educational context of the target language. A German teacher will therefore need to be able to cope with the German school situation in English, and a German child may want to verbalise what happens in his own classroom, rather than in an imaginary one located in an English-speaking country. While no doubt many processes and procedures are identical or at least fairly similar in foreign language teaching situations anywhere, it is important to realise that this is by no means the case for all aspects of the classroom situation, many of which are in fact rather specific to the school setting in a particular country. The Klassenarbeitshefte need to be collected, Peter cannot find his Hausheft, and he has not kept his Aufgabenheft either, the Klassenlehrer wants the Klassenbuch, but the Klassenbuchführer is absent, the Tafeldienst has not cleaned the board, the Klassensprecher is to go to a meeting of the Schülerrat, the pupils want to know whether the next day will be hitzefrei: all of these represent definite concepts in the minds of German teachers and pupils, but have no direct equivalents in the English school setting, and no degree of familiarity with English classroom language in England - with concepts such as prefect, monitor, housemaster, register, rough book - fair (copy) book, etc. - will solve the problem of expressing these German concepts in English.

Clearly, in both cases the problem arises because the concepts suggested by the German language or by the German classroom situation can often only be paraphrased, rather than expressed directly, in a foreign language such as English, where lexicalisations for these concepts may not be readily available. If we accept that it is desirable in principle to cover as much of the actual classroom situation as possible in the foreign language, serious attempts must be made to arrive at acceptable English paraphrases of what German teachers and pupils might want to refer to.

In order to avoid having to invent our own paraphrases for this purpose - with the inherent danger of creating a kind of English only comprehensible within German classrooms - or having to rely on a few individual opinions only - with the danger of overgeneralising usages which may be unnecessarily regional or even idiosyncratic -, 3o native

speakers of English from various parts of the English-speaking world were asked to paraphrase a number of such concepts as they might be needed in German classrooms. In what follows we shall report on this inquiry (2).

3 Method

The inquiry was conducted on the basis of an extensive questionnaire accompanied in each case by a face-to-face interview.

The subjects were native speakers of English from England(=13), the United States(=8), Ireland(=4), Canada(=3), Scotland(=1) and Wales (=1)(N = 3o). All of them happened to be in Germany at the time of the interview, some on a passing visit to the country, some with up to 19 years of residence in it. Most of them were in the age range of 21-3o (=12), the next biggest age group were the 41-5o year-olds(=8), with the other ranges fairly evenly represented (31-4o: =4, under 2o and over 5o:=3each). The large majority must be considered to be very familiar with educational settings: 27 of them were either students (=1o), teachers(=9) or lecturers(=7), and one was a school girl from England. While there is no doubt that the availability - and readiness to co-operate - of subjects for this inquiry was largely dictated by chance, the breakdown of the figures shows that the population was fairly well spread generally, with a desirable overrepresentation of interviewees with explicit educational backgrounds (3).

The questionnaire used in the inquiry was 12 pages long, with over 1oo items arranged in groups pertaining to (i) the personal background of the interviewees, (ii) pupils' equipment, (iii) books, (iv) classroom surroundings, (v) classroom business, (vi) marks, (vii) discipline, and (viii) characterisations of teachers and pupils. The main problem in devising the format of the questionnaire was the need to develop an elicitation procedure enabling subjects who might have no or little command of German and practically no knowledge of German classrooms to co-operate in finding ways of talking about the German classroom situation. Although for each item the starting point naturally was a German concept, it was obviously not possible to simply list these in German and then ask subjects how they would render these in English, e.g. in a format such as "How would you say Die Tafel blendet in English?" Rather, it was necessary to develop descriptions and explanations of these concepts in English, and attempts had to be made to formulate these such as to eventually elicit a paraphrase of the concept required, as e.g. in the following example,

> Some blackboards have a fairly shiny surface, and sometimes the light is reflected in such a way as to make it impossible to read what is on the board from some positions in the room. In such cases, a pupil might complain by saying_____ _____ (questionnaire page 6).

This approach made the questionnaire rather long - 12 pages are quite an imposition (although motivation to complete the task was helped along a little by being able to pay a small remuneration for completion). More importantly, however, there was of course no guaran-

tee that the descriptions were unambiguous enough for the native speakers of English to form an idea of what exactly was meant in each instance. Therefore, interviews were conducted with the subjects after they had made a first attempt to complete the questionnaires on their own, and in these sessions, which took about an hour each, the questionnaires were gone through together and explanations added wherever it appeared that the descriptions had failed to function properly. Typically, subjects had considerable difficulties in switching over to conceptualising matters in the German way - often, initial replies drew attention to the fact that this "did not apply" followed by extensive explanations of the ways in which things were done in their home settings. These explanations were no doubt very interesting in themselves, but provided no help for the problems pinpointed here. A large proportion of the time of the interviews was therefore spent in persuading the subjects to at least try to suggest ways of what one might say for a particular concept (such as Klassenbuch) even if things were (obviously) organised differently at home. It might be added here that the interviews, certainly as far as conducted by the author himself, were personally very satisfying and provided him with highly interesting insights into the details of the day-to-day running of foreign language classes in many parts of the English-speaking world!

4 Results

Only a few of the over 3ooo individual data obtained in this way can be presented and discussed here (cf. note 2).

As a start, it may be instructive to look at the results obtained in reply to descriptions of the German concepts of Lektüre, Notizbuch (of the teacher), Hausheft, Aufgabenheft, Arbeitsheft, and Klassenbuch, which are grouped together here because they all reflect aspects of the Buch/Heft notion in German. The formulation of the items is shown in Table 1, the results are listed separately for each item in Tables 2 - 7.

The results obtained in reply to item 1 (cf. Table 2) show (i) that native speakers would probably avoid the problem by using the title of the booklet to refer to it, and (ii) that if they wanted to be very precise they would have to say something like supplementary text/reading book/reader which is quite a mouthful and therefore unlikely to be used in a normal classroom. Vaguer terms suggested as suitable are reader and reading-book: although the first appears more frequently in the replies, it is often associated with publications providing selections from various sources rather than one story only, so that we would recommend the second as the better term to refer to what is called a Lektüre in foreign language teaching, in Germany - unless one adopts the native speaker's strategy of avoiding the term altogether in the way indicated above.

Table 1

Item Group "Books" (page 3)

1	A book or booklet containing a continuous story, used in foreign language classes as reading material in addition to the actual course book, would be called a
2	The (octavo-sized) book which the teacher carries around with him and uses to note down information about the pupils performance and reminders about work still to be handed in etc. would be called a
3	A book used by pupils for writing in both in class (for classroom work) and at home (for homework) would be called a
4	A book used by pupils to note down what homework has been set for which subject and for which day would be called a
5	A book which the pupils use for writing in only when they do an official test in class would be called a
6	The book which contains the official record of the work done in class, of the marks given for the tests written during the school year, of instances of absence and of arriving late, and of any disciplinary measures taken in the case of individual pupils is called the
	of which there is one per class constantly kept at the school.

Table 2

Item 1 "Lektüre": Results

novel/text	novel/reader	(referred to by title of book)
reading book	reader	supplementary reader
(reading) book/ title would be named	text/story book	reader/reading book
reader	supplementary text	reader/book/story-book/ novel
supplementary reading book	text	reader
story (reading) book	supplementary reading book	novel
reader (more than 1 selection) reading book	literature book	reading book
booklets	reader	reader
text	story-book	
	reader	

Table 3

Item 2 "Notizbuch": Results

the black book	little red book	mark book
teacher's notebook	black book	(scholars') record book
punishment book	grade book	
little book	teacher's record	notebook/diary/mark book
notebook	teacher's notebook	notebook/pocket-book
record book/ teacher's notebook	teacher's handbook the black book	teacher's mark book
grade book/ teacher's notebook	diary/notebook	scholars' record book
notebook	detention book	notebook
notebook/grade book	teacher's notebook	black book
attendance book	mark book/ lesson notebook	diary

Part of the significance attached to the teacher's Notizbuch in the German system stems from its place in the record-keeping, large proportions of which are done at home by teachers, as a consequence of the halfday schooling system customary in Germany: since all official records must remain available at the school, some double book-keeping is necessary to enable teachers to work at home. Thus, although the teacher's Notizbuch is his own private property, pupils recognize immediately that he means business when starting to get his Notizbuch out during a lesson. The various versions offered in reply to this item (cf. Table 3) reflect different aspects of the functions the book may have without, however, covering them all. It seems that a general term such as (teacher's) notebook is the most suitable solution here, with the specification being left to the situational context in class.

A similar solution suggests itself in connection with the Aufgabenheft (cf. Table 4), for which (homework) notebook can serve as a version sufficiently unambiguous in the contexts in which it would be used. More precise versions could be assignment notebook or homework diary (the first probably marked as American English). It should be noted that the Aufgabenheft - which is often a specially arranged octavo-sized booklet with spaces on the margins to enter the lessons of the week in groups following the individual days and with empty lines to write the homework due for these periods against them - plays an important role in the life of a German school child. The half-day schooling usual in Germany tends to relegate most written work to homework, and to keep trace of what is due for when can be quite complicated. Pupils are even asked to show their homework notebooks to see whether they in fact entered the homework set in the right space - and if this happens during an English lesson it would be useful to have an English term for it.

Table 4

Item 4 "Aufgabenheft":Results

note book	assignment book	homework book
homework diary	assignment notebook	notebook
homework-notebook	assignment notepad/ reminder pad	notebook/homework book
notebook/jotter	homework book	notebook
pad/assignment book	assignment book	homework notebook/ jotter
appointment (calendar)book	homework notebook/ diary	exercise book
(homework)assignment notebook	homework diary	homework book/ prep - diary
calendar	homework diary	prep book
notebook reminder	homework diary	prep book
notebook	(write in back of homework book)	homework diary

The results of items 3 (Hausheft) and 5 (Arbeitsheft) are interesting to compare (cf. Table 5) because the distinction reflected in the German terms - the first as the book to be written into in class and at home generally and the second as the book only to be written into for written classroom tests which are thus collected in it - is not identical with distinctions such as rough(copy)book/jotter - fair (copy)book, and a test book is of course usually a (pre-printed) set of papers outlining the questions which are to be answered on individual sheets, for one individual test only. Not surprisingly, the answers show a wide spread of suggestions. Although notebook is technically correct, it should perhaps be avoided for Hausheft if it is made to serve other meanings as well (see above). Jotter/rough book are not quite appropriate for the homework aspect of the use the Hausheft (independantly of what some of the books actually might look like). Thus, exercise book appears as the most convenient solution. For Arbeitsheft, blue book is probably too regional to be useful, and test notebook although technically precise may be too long. This leaves test book in spite of the slightly different associations connected with this term by native speakers of English.

The last item of this group (Klassenbuch, cf. Table 6) is again characterised by the problem of having to find a general term where in English more specific ones are used (register, record files, course record) to refer to the various aspects of the record-keeping needed for the effective administration of school life. It goes without saying that these various aspects are also handled in physically different records, in England - hence the different names - whereas in the German system all the information about the members of a class, the marks received, lessons held, attendances etc. is collected in

Table 5

Items 3/5 "Hausheft" "Arbeitsheft": Results

	Hausheft	
exercise book	notebook	exercise book
jotter	notebook	rough book
copy/exercise book	notebook	exercise book
exercise book	notebook	exercise book
notebook	exercise book	exercise book
notebook	exercise book	exercise book
homework/ class notebook	exercise book (rough book)	exercise book
notebook	exercise book	
notebook	exercise book	
essay book	exercise book	
	Arbeitsheft	
exercise book	blue book	test book
good copybook	blue book	exercise book/test book
test copy	blue book	exercise book/test book
exam book	testbook(let)/ testsheet/exam-	test book
test notebook	test/exam folder	test book
test booklet	test book	test-book
test notebook	test book	(test book)
test booklets	test book	test book
blue book		

Table 6

Item 6 "Klassenbuch": Results

record book	grading book/register	register
report book	roll book register	class register (names only)
assessment records	course record	register
register/class record book	student record (markbook)	(record files)
role book/record book	markbook/teacher's register	official scholars' record book
grade book/official record book	official record book	mark book/register/ punishment book
class book/class register	form book	mark book/register
grade book	black book	record of work book/ lesson planners/ mark book
attendance book	register	
attendance book	(various filing systems)	
	report book	

one book. The answers obtained reflect this difficulty in that they either highlight one of the aspects only (grade book, attendance book, black book etc.), or suggest a combination of these (e.g. mark book/ register/ punishment book)which is descriptive rather than a usable solution, or simply point out the incompatibility of the systems (e.g. various filing systems; class register:names only) which is factually correct but does not solve the problem.

Naturally, it is inappropriate to expect a single "correct" solution, in these cases. What is needed are suggestions which are least likely to be either misleading or incomprehensible to native speakers of the language. Since the term record can be taken to cover records of all kinds (attendance, performance, behaviour, course work etc.) we would suggest record book as a way of referring to the Klassenbuch of German schools, in preference to the clumsier if more precise register and record book put forward in some of the answers.

The second group of items to be presented here tries to elicit paraphrases for the German concepts of Klassenlehrer, Entschuldigung, schwänzen (lesson, day(s)). While the first of these might again come up against the problem of different organisatory structures, the others were included because we were interested in seeing what could - and would - be said colloquially, below the level of official language.

Table 7 shows how the descriptions of the above concepts were formulated.

Table 7

Item Group "Classroom Business" (page 7, items 1-4)

1	In each class, there is usually one teacher who is specially responsible for administrative matters, writing out the reports, checking the register and record book, contacting parents if necessary, organising outings for the class etc.. This teacher is called the _____
2	After an absence a child is supposed to bring _____ _____ from the parents to explain why it was absent.
3	If a child deliberately stays away from an individual lesson without a reasonable excuse s/he is said to _____.
4	If the child deliberately stays away from school altogether s/he is said to _____.

Table 8

Item 1 "Klassenlehrer(in)": Results

Form Master/Mistress	homeroom teacher	class teacher/form teacher
form teacher	homeroom teacher	class(form)teacher
class/form Mistress/ Master	homeroom teacher	form mistress
class teacher	homeroom teacher	form master/ Mistress /Teacher
homeroom teacher	home teacher	form tutor
class sponsor/homeroom teacher	form teacher	class teacher
homeroom teacher	form teacher	form Teacher
homeroom teacher	form teacher	
homeroom teacher	tutor	
home-room teacher/ class administrator	Form Master/Mistress	
	Form Teacher	
	class teacher/form master	

The spread obtained in the replies to the item Klassenlehrer(in) (cf. Table 8) is not due to any lack of a general term - as was the case with some of the items in the previous group - but appears to be the result of straightforward regional variation: our American/ Canadian informants suggested homeroom teacher, our British ones form teacher. In spite of considerable technical differences there seems to

be enough common ground in the function involved to recommend to German teachers the use of either of these terms to express the concept of Klassenlehrer, and the choice would probably depend on the emphasis on either British or American English in that classroom. German teachers might note that the term class teacher was only suggested by 5 (British) informants, all over 4o years of age, 4 of whom with long periods of residence in Germany.

Table 9

Item 2 "Entschuldigung": Results

a note	a written excuse	a note
a note	a letter/note	a sick note/a letter
a letter/note of excuse	a note/note of absence	an absence note/ sick note
an excuse/letter of excuse	a note	a letter
an excuse/an absence note	an explanatory note	a sick note
an excuse	absence note	a letter
an excuse	an absence note	a note
an excuse note	a note	an absence note
excuse	a note	
an excuse	a sick note	
	a letter	
	a note	

The concept of an "Entschuldigung" in the sense of a written statement by parents to explain why their child did not attend school does not seem to be difficult to express (cf. Table 9) and there is quite a range of acceptable terms available, of which the handiest appear to be excuse, (sick/absence)note.

What is perhaps more interesting to see is the way in which staying away from school deliberately, for a lesson or for longer, is referred to informally, i.e. between pupils, or perhaps at home, rather than officially at school. As Table 1o shows, the answers represent both formal and informal versions, with playing truant the most general one in use. Skipping/cutting class is frequently mentioned with the meaning of missing a lesson. Regional variations add colour to the repertoire (dossing, skiving, mitching, wagging it, legging it, playing hookie (AE)) but should perhaps be used only with care (if at all) by non-native speakers. Our initial idea that it might be necessary in English to observe a clear terminological distinction between staying away for an individual lesson, on the one hand, and a day or more on the other, was not systematically supported by the answers obtained: the first case is usually specified by referring to a class or lesson missed that way,and the distinction can therefore be neg-

Table 1o

Items 3,4 "schwänzen" (lesson, day): Results

"schwänzen" (Stunde)		
be dobbing/skiving/ mitching	have skipped class	truant/wag it
skip class	skipped class/cut class	skive off
doss a class	be skipping class	have absconded
mitch/miss a class	be skipping/playing hooky	play truant
cutting a class	skip class	cut the lesson (class)
skip class/play truant	skip class	play truant
be playing hookie/ skipping	skip a lesson/ to skive off	missing a lesson
skip class	mitch/skive	wag
cut class	skive	skiving
		skiving

"schwänzen" (Tag)		
be dobbing	be playing hookey	have left
play truant/scheming school mitching/ skiving	be truant	play truant
play truant/doss	be skipping	play truant
mitch/play truant	skip school/play truant	playing truant
be truant/playing hookey	truant	wag/play truant
delinquent	play truant	play truant
hookie (whole day) skipping (for either)	leg it (Lincolnshire)/ mitch/truant (be truanting)	playing truant
be skipping class/ playing hookey	skive	play truant/mitch take French leave
play hooky	truant/wag it	wagging it
be truant	truant/have the day off	
	have played truant	

lected from a language point of view as not being lexicalised - as is also true for German.

In the final groups of items to be presented and discussed here, we were interested in possible paraphrases of the marks given in the German school system. These go from 1 at the top to 6 at the bottom of the scale, with conventionalised paraphrases (sehr gut, gut, befriedigend, ausreichend, mangelhaft, ungenügend) in addition to the figure names themselves. The descriptions were formulated as in

Table 11, the results are shown in Tables 12a and 12b.

Table 11

Item Group "Marks" (page 9, items 1 - 6)

There are six marks available in the German system to assess the child's school performance. They go from one to six. The first four are pass marks and can be paraphrased as follows

the highest possible mark is a "1" = _____
the second highest mark is a "2" = _____
the neutral, average mark is a "3" = _____
the "just pass" mark is a "4" = _____

The other two marks are non-pass marks and express two levels of failure,

inadequate work but not beyond help is a "5" = _____
the lowest possible mark is a "6" = _____

The results obtained are not easy to interpret and are reminders of the linguistic truism that the meaning of items in a closed set is largely determined by the co-presence, and rank, of the other items in it. While most informants agreed on using excellent for the very best mark, very good can only be the very best if there is no excellent to precede it: in all other cases it is used to refer to the second best mark. A similar shift can be observed in connection with good: it is only used for second best, if the top mark is very good, or excellent with no very good following: in all other cases it refers to the average mark. A frequent alternative given for the average mark is satisfactory, which, however, also appears as a mark a grade lower than average, in other words as the lowest pass mark. The two non-pass marks are either both given as fail, or a distinction is drawn between poor and very poor/bad. It is interesting to observe that many native speakers tried to express German marks in terms of their own marking conventions (using letters from A-E/F, or giving percentages); on the other hand, practically all suggestions were consistent in themselves even if the same term (e.g. satisfactory) would serve different purposes in the different schemes suggested.

On the basis of the answers obtained here, we would recommend to the German teacher of English the use of the following terms - as alternatives to the names of the figures as such -: 1 = excellent/very good, 2 = good, 3 = satisfactory/average, 4 = pass/fair, 5 = poor, 6 = very poor/bad.

Table 12a

Marks 1-3: Results

"sehr gut"	"gut"	"befriedigend"
A (75%) excellent	B (55%) very good	C (45%) good
excellent A	very good B	good C
A (85-1oo%) excellent	B (7o-84%) very good	C (55-7o%) good
excellent	very good	good/fair
excellent work	good work	satisfactory
A excellent	B good	C average
A very good/ excellent	B good	C satisfactory
A (excellent)	B above average	C average
A excellent	B above average	C average
A excellent	B good	C average
A excellent	B above average	C average
A excellent	B good	C satisfactory
9o-1oo% A exceptional	75-9o% B good	6o-75% C average
A excellent	B very good	C+, C, C-good
1oo% A the best	8o% B above average	65% C average
excellent	good	average
excellent	very good	fairly good
A excellent	B very good	C satisfactory
A excellent	B good	C average
very good	good	average
A excellent	B very good	C good
very good/ excellent	good	satisfactory
excellent	good (very good)	average/good
A	B	C
1o/1o topmark	9/1o	5/1o-4/1o
A excellent	B very good	C good
very good	good	satisfactory
A excellent	good	average

Table 12b

Marks 4-6: Results

"ausreichend"	"mangelhaft"	"ungenügend"
C/D (35%) average below average	D (35%) below average	E/F (25%) poor failure
fair D	bad E	very bad V
D (4o-54%) satisfactory	E (35-39%) insufficient	F (1o-24%) N.G (no grade)
a pass/a bare pass	fail	a hopeless fail
passing	unsatisfactory	failing "a zero"
D below average	I incomplete	F failure
D sufficient	F fail/inadequate	F fully inadequate
D below average	possibility of failing	flunk
D unsatisfactory	F fail	F fail
D adequate	inadequate	F failure
D below average	F failed	F failed
D passing	E failure with hope	F failure
5o-6o% D sufficient	4o-5o% E inadequate	below 4o% F wholly inadequate
D satisfactory	E insatisfactory	F definite fail
5o% D below average	2o-5o% E	o-2o% F failure
fair	poor	very poor
fair	below average	poor
D could do better	E	V unclassified
D satisfactory	E below average	F very poor
satisfactory	below average	very poor
D fair	E poor	F very poor
fair	poor/weak	very poor/weak
pass	poor/weak	very poor
D 4/1o	E 3/1o	F o/1o fail/worst mark
D fair	E poor	F very poor
fairly good	only fair	weak E
below average	poor	very bad

5 Conclusions

In the present paper an attempt has been made to present and discuss some of the results obtained in the course of an investigation into the language which native speakers of English would use to refer to selected aspects of the German classroom situation.

We have argued that it is important to include the actual classroom situation of the learners as far as possible within the foreign language parts of foreign language lessons. However, foreign languages are typically learnt in the educational environment of the learner's home contry, rather than that of the target language. This means that a direct transfer of the classroom language connected with e.g. English institutional surroundings is not possible: German children and teachers need to find ways of verbalising their own surroundings which very often are rather different.

In order to avoid making suggestions which would lead to a negative kind of "classroom English" incomprehensible to native speakers of the language, outside German classrooms, a survey was conducted asking 3o native speakers of English from various parts of the English-speaking world to paraphrase in English a number of classroom situations as they might arise in German classrooms. This was done on the basis of an extensive questionnaire, followed up by a face-to-face interview. A major problem turned out to be the need, for the questionnaire, to develop an elicitation procedure enabling also subjects with no or next to no German and practically no knowledge of German classrooms to participate in finding ways of talking about the German classroom situation. The solution eventually adopted of presenting the subjects with carefully phrased descriptions (in English) of the various concepts to be expressed hopefully in usable English paraphrases was not always successful: major parts of the interviews were spent on additional explanations where care was needed not to suggest what the interviewer wanted to hear in the first place.

The results - only some of which could be presented and discussed here - show that the procedure adopted for this investigation is a possible way of arriving at the information required. As we have seen, this does not mean that the answers obtained are always easy to interpret nor that can they be incorporated into e.g. a German English teacher's classroom language repertoire directly, i.e. without any further interpretation. But we hope to have been able to show that the data obtained in this way are both necessary and useful as a basis for recommendations in this field.

As stated at the beginning, this investigation is part of a larger project concerned with the development of teaching materials for the field of "classroom language" for German teachers of English. It has proved to be very helpful in placing these materials on sounder foundations and will hopefully, in the long run, make a substantial contribution to encouraging teachers and pupils in Germany to exploit more of their own classroom situation for foreign language interaction than is often the case at present, by showing how this can actually be done.

Notes

1 supported by grant No. 8546 of the University of Bielefeld.

2 Only some of the data obtained in the course of this investigation can be presented here. More information is available from the author, on request. A fuller report is in preparation.

3 I wish to record my thanks here to John Hardman, who as a student assistant to the project helped to find a large number of the native speakers who eventually participated, and who conducted a fair number of the interviews.

Bibliography

Hughes, G.S. 1981. A handbook of classroom English. London: Oxford University Press.

Voss, B. 1984. "Classroom language - A neglected area in foreign language teaching and testing". In Culhane, T., Klein-Braley, C., and D.K. Stevenson (eds.). 1984. Practice and Problems in Language Testing. Colchester: University of Essex, Department of Linguistics, Occasional Papers No. 29: 177-193.

AFinLA Series

The purpose of AFinLA is to disseminate information on research into applied linguistics in this country and abroad through publications, seminars and conferences. AFinLA is the Finnish affiliate in AILA (Association Internationale de Linguistique Appliquée).

The current membership fee is FIM 50 (student fee FIM 25). Information on AFinLA activities appears in the Kielikeskusuutisia (News Bulletin of the Language Centre for Finnish Universities). Members receive the AFinLA Yearbook and are entitled to a 50 per cent discount on AFinLA publications.

Members are requested to order publictions from the Department of English, University of Jyväskylä. Non-members should send their orders to Akateeminen kirjakauppa, Helsinki (see back cover).

Recent publications:

N:o 28. Kari Sajavaara, Anne Räsänen & Tuula Hirvonen (toim.), AFinLA:n vuosikirja 1980. Jyväskylä 1980. (165 s., 30 mk)

N:o 29. Matti Luukkainen und Konrad Schröder, Diversifikation und Motivation im Fremdsprachenunterricht. Turku 1981. (101 s., 25 mk)

N:o 30. Kari Sajavaara & Matti Leiwo (toim.), Psykolingvistisiä kirjoituksia I. Jyväskylä 1980. (119 s., 25 mk)

N:o 31. Matti Leiwo & Kari Sajavaara (toim.), Psykolingvistisiä kirjoituksia II. Jyväskylä 1981. (130 s., 30 mk)

N:o 32. Kari Sajavaara & Anne Räsänen (toim.), AFinLA:n vuosikirja 1981: Tätä mieltä kielentutkimuksesta ja -opetuksesta. AFinLA:n 10-vuotisjuhlasymposiumin 14.-15.11.80 puheenvuorot. Jyväskylä 1981. (105 s., 25 mk)

N:o 33. Jorma Tommola & Raija Ruusuvuori (toim.), AFinLA:n vuosikirja 1982: Näkökulmia kirjoitettuun kieleen. Turku 1981. (126 s., 30 mk)

N:o 34. Kari Sajavaara, Maija Kalin & Matti Leiwo (toim.), Psykolingvistisiä kirjoituksia III. Jyväskylä 1982. (220 s., 40 mk)

N:o 35. Seija Tiisala (utg.), Grammatiska kontraster. Helsingfors 1983. (211 s., 45 mk)

N:o 36. Heikki Nyyssönen & Anna Mauranen (toim.), AFinLA:n vuosikirja 1983. Oulu ja Helsinki 1983. (168 s., 35 mk).

N:o 37. Kari Sajavaara, Jorma Tommola & Matti Leiwo (toim.), Psykolingvistisiä kirjoituksia IV. Jyväskylä 1984. (122 s., 35 mk).

N:o 38. Kari Sajavaara, Matti Leiwo & Mervi Eloranta (toim.), AFinLA:n vuosikirja 1984. Jyväskylä 1984. (122 s., 35 mk).

N:o 39. Kari Sajavaara, Jorma Tommola ja Seija Xystö (toim.), Psykolingvistisiä kirjoituksia V. Jyväskylä 1985 (128 s., 35 mk).

N:o 40. Viljo Kohonen, Hilkka von Essen and Christine Klein-Braley (eds.), Practice and Problems in Language Testing 8. Tampere 1985. (218 pages, FIM 60,oo).

SUOMEN SOVELTAVAN KIELITIETEEN YHDISTYKSEN
(AFinLA) JULKAISUJA

PUBLICATIONS DE L'ASSOCIATION FINLANDAISE
DE LINGUISTIQUE APPLIQUÉE (AFinLA)

Jakelu:
Distributor:

ENGLANNIN KIELEN LAITOS
JYVÄSKYLÄN YLIOPISTO
40100 JYVÄSKYLÄ

AFinLA:n jäsenet
AFinLA members

AKATEEMINEN KIRJAKAUPPA
POSTILOKERO 128
00101 HELSINKI

Muut kuin AFinLA:n jäsenet
Others

ISBN 951-9388-27-3
ISSN 0781-0318

216